Helsinki University of Technology - Publications in Geoinformatics and Cartography
Teknillisen korkeakoulun geoinformatiikan ja kartografian julkaisuja
Espoo 2009 TKK-GC-9

# DISCOVERING SPATIO-TEMPORAL RELATIONSHIPS: A CASE STUDY OF RISK MODELLING OF DOMESTIC FIRES

Olga Špatenková

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Engineering and Architecture for public examination and debate in the Auditorium M1 at Helsinki University of Technology (Espoo, Finland) on the 11$^{\text{th}}$ of December 2009, at 12 noon.

Helsinki University of Technology
Department of Surveying

Teknillinen korkeakoulu
Maanmittaustieteiden laitos

Distribution:
Helsinki University of Technology
Department of Surveying
P.O.Box 1200
Tel. +358 9 47023911
Fax. +358 9 465077
E-mail: ritva.laitikas tkk.fi

© Olga Špatenková

Maps on the front cover:
Changing density of domestic fires in Helsinki during the different periods of the day.

ISBN 978-952-248-191-171 (printed)
ISBN 978-952-248-233-4 (pdf)
ISSN 1795-5432

Otamedia Oy
Espoo 2009

# Abstract

A systematic risk analysis for mitigation purposes plays a crucial role in the context of emergency management in modern societies. It supports the planning of the general preparedness of the rescue forces and thus enhances public safety. This study applies the principles of knowledge discovery and data mining to support the development of a risk model for fire and rescue services. Domestic fires, which are a serious threat in an urban environment, are selected to demonstrate the methods.

The aim of the research is to identify important factors that contribute to the probability of the occurrence of domestic fires. Various physical and socio-economic conditions in the background environment are analysed to provide an insight into the distribution of domestic fires in relation to underlying factors.

Following the cross-disciplinary nature of data mining, this study offers a set of distinct methods that share the same goal – to identify patterns and relationships in data. The methods originate in different scientific fields, such as information visualisation, statistics, or artificial intelligence. Each of them reveals different aspects of the existing relations, which supports an understanding of the phenomenon and thus expands the expert knowledge.

The application of data mining techniques is not straightforward because of the specific nature of geospatial data. This study documents the analysis process in order to provide guidelines for potential future users. It considers the suitability of the methods to handle spatial and spatio-temporal data with special attention to the GIS-motivated conceptualisation of the problem being analysed. Furthermore, the requirements for the user to be able to apply the methods successfully are discussed, as is the available software support.

# Preface

My relation with Finland started during my master's studies as short tentative exchange stay in 2001. But Finland is a trap, which does not allow one to leave easily. So, after graduating from Helsinki University of Technology, I have also started my doctoral studies there. I am grateful for this chance and experience I have gained.

My wonderful supervisor Prof. Kirsi Virrantaus is the key person behind that all. She gave me opportunity to work in her lab and carry out the research. I would like to express my gratitude for her immense support and encouragement during my studies in Finland.

I also appreciate all the financial support I have received during my doctoral studies. The researcher training scholarship from Helsinki University of Technology allowed me to complete the theoretical part of my studies at the beginning. The Finnish Fire Protection Fund of the Ministry of the Interior and the Academy of Finland have been funding my further research.

A part of my research originates from my visit to the ITC International Institute of Geoinformation Science and Earth Observation. I am thankful to the ITC and Prof. Alfred Stein to facilitate this. Prof. Alfred Stein instructed my work and I want to thank him for his patient guidance and encouragement.

I also want to thank Prof. Stewart Fotheringham and Dr. Kati Tillander, who were the preliminary examiners of this thesis, for their constructive comments and suggestions for improvements.

I am grateful to all my colleagues from the lab for a unique working environment. I have also enjoyed friendly atmosphere of the lunch and coffee times. I would also like to thank Urška Demšar, who cooperated with us in the research intensively.

Many thanks go to my friends, who cheered me up outside the working place. Special thanks belong to my friends I met in Finland, who willingly offered me five-

star services during my frequent visits.

I would like to express my sincere thanks to my parents and other members of my family for all their love and support. Finally, my greatest thanks belong to my husband Petr for his patience, encouragement and looking after me not only during the long months of writing my thesis.

# Contents


**1 Introduction** **1**

1.1 Motivation . . . . . 1

1.2 Big picture . . . . . 2

1.3 Aim and objectives . . . . . 6

1.4 Research questions . . . . . 8

1.5 Limitations . . . . . 10

**2 Geographical data mining** **12**

2.1 Data mining and knowledge discovery . . . . . 12

2.2 Mining of geospatial data . . . . . 16

2.3 Representation of knowledge discovered . . . . . 19

**3 Description of case application and materials** **23**

3.1 Domestic fires . . . . . 23

3.2 Data description . . . . . 24

3.2.1 Incident records . . . . . 24

3.2.2 Building characteristics . . . . . 25

3.2.3 Census data . . . . . 27

**4 Visual data mining** **33**

4.1 Method . . . . . 33

4.2 Data pre-processing . . . . . 35

4.3 Results . . . . . 36

4.4 Conclusions . . . . . 40

4.4.1 Capturing spatial and temporal aspects . . . . . 40

4.4.2 Type of knowledge discovered . . . . . 43

4.4.3 Weaknesses . . . 44
4.4.4 Requirements for the user . . . 44

**5 Contingency tables 46**
5.1 Method . . . 46
5.2 Data pre-processing . . . 48
5.3 Results . . . 49
5.4 Conclusions . . . 51
5.4.1 Capturing spatial and temporal aspects . . . 51
5.4.2 Type of knowledge discovered . . . 53
5.4.3 Weaknesses . . . 54
5.4.4 Requirements for the user . . . 54

**6 Point pattern analysis 56**
6.1 Method . . . 56
6.2 Data pre-processing . . . 60
6.3 Results . . . 60
6.4 Conclusions . . . 67
6.4.1 Capturing spatial and temporal aspects . . . 67
6.4.2 Type of knowledge discovered . . . 69
6.4.3 Weaknesses . . . 69
6.4.4 Requirements for the user . . . 70

**7 Geographically weighted regression 71**
7.1 Method . . . 71
7.2 Data pre-processing . . . 75
7.3 Results . . . 75
7.4 Conclusions . . . 82
7.4.1 Capturing spatial and temporal aspects . . . 82
7.4.2 Type of knowledge discovered . . . 90
7.4.3 Weaknesses . . . 91
7.4.4 Requirements for the user . . . 91

**8 Self-Organising Maps 92**
8.1 Method . . . 92

8.2 Data pre-processing . . . . . 95
8.3 Results . . . . . 96
8.4 Conclusions . . . . . 103
8.4.1 Capturing spatial and temporal aspects . . . . . 103
8.4.2 Type of knowledge discovered . . . . . 103
8.4.3 Weaknesses . . . . . 103
8.4.4 Requirements for the user . . . . . 104

**9 Discussion** **106**

**10 Conclusions** **110**
10.1 Conceptualisation . . . . . 110
10.2 Methods . . . . . 112
10.3 Implications for risk modelling . . . . . 116
10.4 Future research . . . . . 118

# List of Figures


1.1 Emergency management framework. . . . . 3
1.2 Relation between elements of risk. . . . . 5
1.3 Framework of the study. . . . . 6

2.1 Data mining process. . . . . 14
2.2 Data, information, knowledge, and wisdom hierarchy. . . . . 20

3.1 Distribution of domestic fires in the study area. . . . . 25
3.2 Average density of domestic fires on different temporal scales. . . . . 26
3.3 Average density of domestic fires in relation to building attributes. . . 29
3.4 Distribution of socio-economic population attributes. . . . . 31
3.5 Average density of domestic fires in relation to socio-economic aspects. 32

4.1 Visual data mining system. . . . . 34
4.2 Distribution of d-fires in relation to e-fires. . . . . 37
4.3 Bivariate matrix for the temporal fire categories. . . . . 38
4.4 Temporal changes in the spatial distribution of domestic fires in relation to population density. . . . . 39
4.5 PCP for high density of domestic fires. . . . . 41
4.6 PCP for density of domestic fires and building age. . . . . 41
4.7 Connection between the PCP and the map. . . . . 42

6.1 Density plots of domestic fires. . . . . 61
6.2 $\hat{K}$-functions for domestic fires. . . . . 62
6.3 $\hat{G}$-functions showing relations between domestic fires and population attributes. . . . . 64

6.4 $\hat{G}$-functions showing relations between domestic fires and building attributes. . . . . 66
6.5 Goodness of fit of the selected best temporal models. . . . . 67
6.6 Fitted density function of the selected best temporal models. . . . . 68

7.1 A spatial weighting function. . . . . 73
7.2 Weighting schemes. . . . . 74
7.3 GWR results for a full model of all fires (1). . . . . 76
7.4 GWR results for a full model of all fires (2). . . . . 79
7.5 GWR results for a full model of all fires (3). . . . . 80
7.6 GWR results for a full model of all fires (4). . . . . 81
7.7 Local $r^2$ for reduced models. . . . . 85
7.8 Standardised residuals for reduced models. . . . . 86
7.9 Parameter values of population density for reduced models. . . . . 87
7.10 Parameter values of density of workplaces for reduced models. . . . . 88
7.11 Parameter values of households with children for reduced models. . . . 89
7.12 Parameter values of households with adults for reduced models. . . . . 90

8.1 Neighbourhood function in a SOM organised in a hexagonal lattice. . 93
8.2 Clusters identified from the distance matrix and data histogram for the incident dataset. . . . . 96
8.3 Location of the identified clusters in the component planes. . . . . 98
8.4 Discovering relationships between the attributes from the component planes. . . . . 99
8.5 Clusters identified from the distance matrix and data histogram for the grid representation. . . . . 100
8.6 Location of the identified clusters for the grid representation of the data in the component planes. . . . . 101
8.7 Identifying differences between e-fires, d-fires, and n-fires in the grid representation from the component planes. . . . . 102

# List of Tables


3.1 Reclassification of building attributes. . . . . 28
3.2 Classification of census data. . . . . 30

5.1 A contingency table. . . . . 47
5.2 Results of the $\chi^2$ test. . . . . 49
5.3 Cramér's $V$ values. . . . . 50
5.4 Detailed view of contingency tables and $\chi^2$ statistics for domestic fires and population density. . . . . 51
5.5 Detailed view of contingency tables and $\chi^2$ statistics for domestic fires and workplace density. . . . . 52
5.6 Detailed view of contingency tables and $\chi^2$ statistics for domestic fires and density of households with adults. . . . . 52

6.1 Average density ($\times 10^{-7}$) of daytime categories of domestic fires per $m^2$ studied for different days of the week. Notice the increased values during Friday and Saturday evenings. . . . . 62
6.2 Parameter values for the selected model (significant parameters marked with *). . . . . 65

7.1 Global regression results for full models. . . . . 77
7.2 GWR results for full models. . . . . 78
7.3 Global regression results for reduced models. . . . . 83
7.4 GWR results for reduced models. . . . . 84

10.1 Summary of the aspects revealed by the methods applied. . . . . 114

# Chapter 1

# Introduction

## 1.1 Motivation

Risk analysis plays a crucial role in the context of emergency management in modern societies, as it supports the mitigation of emergencies and the planning of the general preparedness of the rescue forces. A suitable allocation of resources on the basis of thorough risk analysis can improve the emergency response of rescue units and thus enhance public safety. The responsible authorities therefore recognise risk analysis as an important part of accident prevention and focus on the development of reliable models reflecting possible threats and the associated risks.

In Finland, a systematic risk analysis forms the official basis for planning the resources for civil protection [Lonka, 1999]. According to the Guidelines on Preparedness of Municipal Fire Brigades issued by the Ministry of the Interior in 1992, the analysis is performed at a municipal level to identify potential threats and evaluate the abilities of fire brigades to overcome them.

The Rescue Office in the city of Espoo has developed a GIS application to be used for the planning of resource allocation in each municipality [Ihamäki, 1997]. The application creates a zone map expressing the required level of preparedness for emergency situations. Risk level 1 areas must be reached in less than 6 mins after the announcement. The response times for areas with risk levels 2, 3, and 4 are 12, 15, and 20 min, respectively. The calculation of the risk zones is based on three explanatory variables: population density, the floor area of buildings, and the probability of traffic accidents.

The current application offers the basic grounds for planning the emergency preparedness; however, it oversimplifies the issue of risk assessment. The model generalises a number of different types of incidents into a single class, which prohibits consideration of the reasons for the different incidents as well as their consequences. The input variables also deserve more attention. For example, the population density is based on the permanent addresses of the inhabitants, which does not reflect temporal variations in their distribution during the daytime. The development of a spatio-temporal population models reflecting the actual distribution of the inhabitants has been proposed as a solution to this problem [Ahola et al., 2007; Krisp, 2008; Molarius et al., 2009].

In addition, Krisp et al. show in their study that the relationship between the occurrence of incidents and a high population density is not as strong as assumed, and hence it is desirable to explore the influence of other input variables [Krisp et al., 2005]. The research group has applied different exploratory methods to unveil the spatio-temporal relationships between incidents and various geographic or socio-economic aspects describing the background environment, e.g. [Krisp and Karasová, 2007; Špatenková et al., 2007; Špatenková and Stein, 2009]. The results support the development of a risk model with increased reliability, enhancing the performance of the rescue services.

## 1.2 Big picture

Emergency management is the discipline and profession of applying science, technology, planning, and management to deal with extreme events that can injure or kill large numbers of people, do extensive damage to property, and disrupt community life [Hoetmer, 1991]. A conceptual framework frequently used for emergency management [Godschalk, 1991; Cova, 1999] is described below. It uses a temporal dimension of the extreme events to form a cycle of four, often overlapping, phases: mitigation, preparedness, response, and recovery (Figure 1.1).

The aim of mitigation is to reduce the potential harm to humans, property, and the environment by applying methods for risk analysis and management. The development of operational capabilities to facilitate an effective response to an emergency situation is performed during preparedness. The response phase concerns measures taken to minimise the damage, which are connected with the occurrence of an emer-

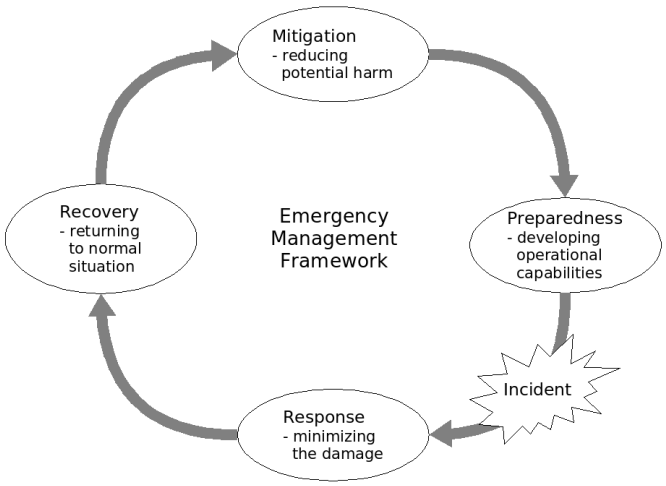



Figure 1.1: Emergency management framework, adapted from [Godschalk, 1991; Cova, 1999].

gency. Finally, a normal situation is returned in the recovery phase.

As the problems inherent in emergency management are spatial, Geographical Information Science (GIS) is adopted in this field to support spatial decision making in each of the phases. The mitigation and preparedness phases deal mainly with risk assessment, impact modelling, and vulnerability analysis. The situation picture and tools that support shared situation awareness are used during the response phase and technologies supporting the regular activities of society serve during the recovery phase [Seppänen and Virrantaus, 2009].

This study aims to support the development of risk models for mitigation purposes. The term 'risk' is understood in a broad sense; however, its implications vary according to the specific environment and application under consideration. It is therefore desirable to specify the meaning of a risk in order to distinguish it from a hazard, threat, or vulnerability and define the related terms.

- **A threat or hazard** is a condition or physical situation with a potential for undesirable consequences to people, property, or the environment, occurring as a result of natural or human activities [Godschalk, 1991; Ayyub, 2003; Modarres, 2006].

- **Consequences** represent the degree of damage or loss from an event of failure [Ayyub, 2003].
- **An incident** is a neutral term for events whose occurrence interrupts the normal situation.
- **A risk** is a measure of a potential hazard [Modarres, 2006]. It represents the probability that a hazard will occur during a particular time period [Godschalk, 1991]. As shown in Figure 1.2, a risk consists of two components: the likelihood of the occurrence of an event and the severity of the impact of that occurrence of an event on the basis of event scenarios [Ayyub, 2003].
- **Vulnerability** affects both the probability of a hazard and its consequences. It is a set of characteristics of the system that creates the potential for harm, but is independent of the probabilistic risk of the occurrence of any particular hazard [Sarewitz et al., 2003; Modarres, 2006]. It can also be interpreted as a susceptibility to injury or damage from hazards [Godschalk, 1991]. Vulnerability results in a weakness that can be exploited by an adversary to cause damage [Ayyub, 2003]. The relation between vulnerability and risk is shown in Figure 1.2.
- **Safety** can be defined as the judgment of risk acceptability. It is a relative term, as the decision depends on the individual making the judgment [Ayyub, 2003].

Risk analysis is the process of characterising, managing, and communicating the risk in order to measure the potential loss and identify the elements that contribute most to such losses. Risk assessment is the first step and the core part of this process that aims to quantify the probabilities and magnitudes of losses resulting from exposures to hazards using formal methods. Risk analysis further includes risk management and risk communication. The potential of the magnitude and the contributors to the risk are estimated, evaluated, and controlled through risk management. Risk communication is the process of sharing information about the nature of the risk, the approach to risk assessment, and the options of risk management between the parties involved [Modarres, 2006].

Generally, the risk assessment aims to answer three basic questions [Ayyub, 2003].

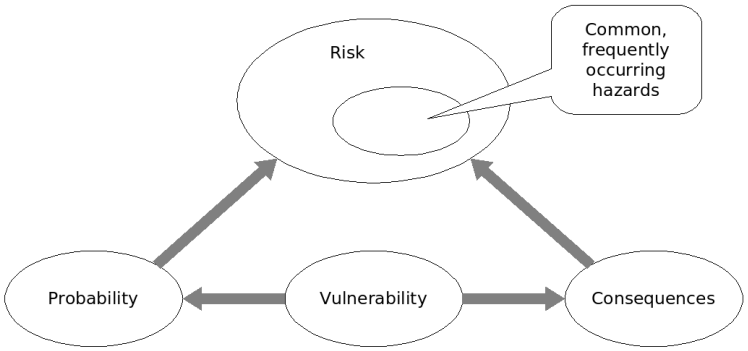



Figure 1.2: Relation between elements of risk.

1. What can go wrong?
2. What is the likelihood that it will go wrong?
3. What are the consequences if it goes wrong?

The identification of potential hazards and scenarios is an important stage of risk assessment in order to describe the risk adequately. It consists of recognising the hazards and defining their characteristics. This stage is highly subjective, as it depends on the experience of the experts involved. For hazards that have been identified, the estimate of the risk is commonly expressed as the product of the probability of the occurrence ($Pr(e)$) and consequences ($C(e)$) of an event ($e$):

$$R(e) = Pr(e) \times C(e). \tag{1.1}$$

Identifying the factors affecting the occurrence of hazards and expressing their influence, which is necessary for the construction of a probability model, is the focus of this study. The complexity of the phenomena underlying the possible threats in space and time makes the task difficult. The use of sophisticated methods is necessary to provide an insight into the data being analysed. The exploration of suitable datasets that cover various physical and socio-economic aspects enables relations that exist in the background environment to be inspected. Their identification and understanding

create knowledge, which can be used to improve the reliability of consecutive risk models.

## 1.3 Aim and objectives

The aim of this research, as illustrated in Figure 1.3, is to offer a cross-disciplinary collection of methods revealing different aspects of the data to support the development of a fire risk model. Although the methods selected for this study originate in different scientific fields, they share the same goal – to identify patterns and relationships in data. Effort is made to combine distinct methods, rather than to remain within a single framework, in order to expand the knowledge that has been discovered. The study focuses specifically on a detailed description of the type of knowledge each method can offer. It also considers the suitability of the methods applied in terms of user-friendly environments and the complexity of their theoretical or computational requirements. At the same time the study documents the process of the discovery of spatio-temporal knowledge, with special attention being paid to distinguishing the roles of domain and GI experts in the interpretation of the results.

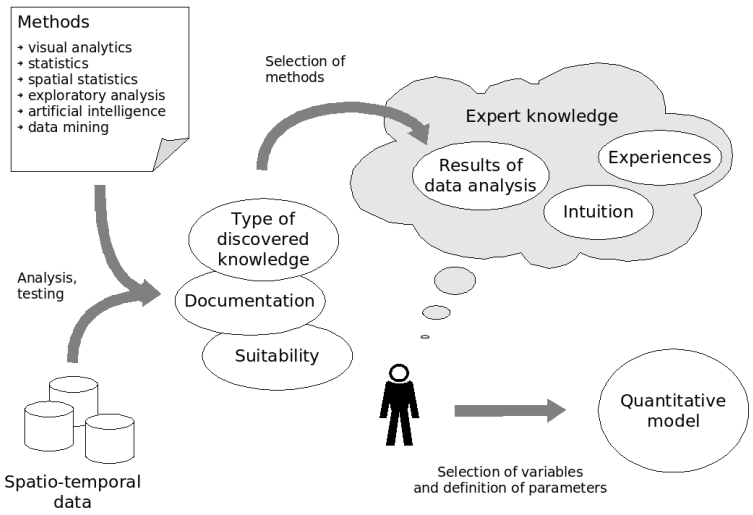



Figure 1.3: Framework of the study.

The hypothesis of this research can be formulated as follows.

*Careful data analysis using several distinct methods reveals different as-*

*pects of the relationships existing in the data. These aspects enhance the understanding of the phenomenon and enable detailed knowledge to be formed, which supports the construction of a realistic risk model.*

Specific research objectives based on the aforementioned ideas are as follows.

- **Application of different methods originating in distinct domains to discover relationships in spatio-temporal data.**

  As geographical data have a specific nature, the application of methods used in other fields (classical non-spatial data analysis and mining methods) is not straightforward. It is important to cover various spatial aspects, such as geometrical and topological relationships, autocorrelation, heterogeneity, temporal variations, and also the question of spatial and temporal resolution. Careful conceptualisation during the data pre-processing step is usually required to cope with this problem.

- **Documentation of the methods.**

  This study aims to provide guidelines for the process of knowledge discovery from spatial and spatio-temporal data. Special attention is therefore paid to the detailed description of the analytical procedures.

- **Analysis of the results.**

  The main focus is on a theoretical point of view, exploring the type of knowledge discovered. Special attention is paid to the outcomes that are related to the management of spatio-temporal dimensions of data and the ability of each method to cope with them.

- **Considering the suitability of the methods for spatio-temporal knowledge discovery.**

  Each of the methods presented demands different volumes of pre-processing, involves a combination of several software tools, and requires an understanding of different amounts of theory. Future users need to consider these aspects in order to select the methods that best suit their needs.

- **Contribution to the research challenges of spatio-temporal data mining.**

Data mining and knowledge discovery in the realm of geospatial data are an open research topic currently. The study contributes to the field by means of test cases, in which it demonstrates spatio-temporal data mining via the specific example of domestic fires in an urban environment.

- **Setting requirements for the development of a spatio-temporal data mining toolbox.**

  On a technical level, this research opened up the issue of the absence of connections between the different available software applications, as well as the variety of user interfaces and visualisation styles used needed to perform the methods. The study can be used to specify the requirements for a data mining toolbox to be implemented on top of existing GI software.

## 1.4 Research questions

Geospatial data mining and knowledge discovery represent an important direction in spatial analysis in a data-rich environment. They collect the work of various groups, including geographers, GI scientists, computer scientists, and statisticians to discover new and unexpected patterns, trends, and relationships hidden in large and diverse geographical databases. Geospatial data mining is a promising but young discipline, facing many research problems, e.g. the consideration of spatial dependency and heterogeneity or the development of techniques for automated and visual mining with a suitable database support [Miller and Han, 2001].

The present study aims to contribute to this field by offering applications from the domain of risk management. Each of the methods applied is analysed in detail in terms of its suitability for geospatial data and user requirements. Special attention is paid to analysing the knowledge each of the methods gains by comparing the similarities and differences between them. The experience gained is used to specify the framework and needs of geospatial data mining more closely.

The study aims to answer the following research questions for each of the methods applied.

- **How does the method capture the spatial and temporal aspects?**

  The changing of data properties over space and time distinguishes geographical information science from other non-spatial domains. The study illustrates the

peculiar issues arising from the application of the method to the distribution of fire incidents and summarises the experiences from a perspective of compatibility with special aspects of geographical data.

- **What is the type of knowledge discovered by the method?**

  Each of the methods reveals the relationships between the data in different terms. This study describes in detail the type of knowledge which can be acquired by the particular method and demonstrates the message by means of a practical example.

- **What are the weaknesses of the method?**

  Each of the methods is fraught with subjective decisions and weaknesses that a user should be made aware of. The study highlights these points for a successful application of the methods.

- **What are the user requirements for the method to be applied?**

  The study aims to provide practical guidelines for future users in terms of available software tools and the necessary pre-processing steps and their complexity. In addition, it indicates the amount of theoretical knowledge and experience required from the user to be competent to apply the method and interpret the results properly.

On top of the particular questions to be answered for each of the selected methods, the study also deals with the following research issues.

- **How can the application of different data mining methods support knowledge discovery?**

  The study provides an alternative approach to spatio-temporal data mining by analysing and comparing different methods. The study considers how such an approach supports the knowledge being discovered. It examines what new knowledge is gained from each method and points out the similarities and differences.

- **How can the knowledge revealed support risk modelling?**

  The motivation of this research is to support the development of the risk model for fire & rescue services. The study discusses the usability of the knowledge

revealed for risk modelling and also the impact of the results on mitigation procedures.

- **What is the role of the GIS expert in the process of knowledge discovery?**

  Knowledge discovery is a complex process connecting experts from various application areas with information scientists via a common goal. Each of the parties involved supports the knowledge discovery from different viewpoints. While domain experts possess knowledge and experience in their specific field, a GIS expert offers a set of tools for advanced analysis related to geographical data. Understanding their roles is necessary for their successful interaction during the process.

- **What are the design requirements for a data mining toolbox?**

  The study points out a drawback in appropriate software support for knowledge discovery within an existing GI platform. On the basis of the experience gained, it suggests the design requirements for the development of a geographical data mining toolbox.

## 1.5 Limitations

As indicated in Figure 1.2, the study deals with common hazards rather than with emergency situations that are unexpected. Domestic fires are used as a case example to represent such hazards. They can be effectively managed using probabilistic risk modelling based on accurate information about the incidence of events.

Risk is commonly viewed as the probability of the occurrence of a hazard and the consequences associated with it. The study focuses on the former, incident probability. As an assessment of consequences requires a different strategy from a probabilistic approach based on collected data, the consequences are left out of consideration.

The study analyses different methods for supporting expert knowledge in the probability distribution of emergency events (see Figure 1.3). As the main focus is on a theoretical point of view, the actual interpretation of the results is limited to statements about the observations gained from the data. Drawing serious conclusions is left for domain experts.

The study emphasises the suitability of the methods for treating geographical data, which is demonstrated by means of a case study. The actual implementation of the probability model, which requires further evaluation of the results, is, however, beyond the scope of this study.

The methods presented in this study are data-driven. The reliability of the results therefore depends on the quality of the original data, which are burdened with uncertainty. As the study focuses on the suitability of the methods rather than on actual results, the uncertainties of the source data are not in the focus of this research.

The knowledge that is discovered in this study is understood as the characterisation of the phenomenon being studied within the background environment. It is represented as important possible influences identified from the analysis. An exact quantification of the relationships, as well as of the uncertainty associated with them, is postponed until the later phase of the construction of the probability model for risk evaluation.

# Chapter 2

# Geographical data mining

## 2.1 Data mining and knowledge discovery

Recent technological progress in methods for data acquisition and storage has resulted in the collection of large amounts of data. Such data-rich collections conceal potentially useful information and knowledge, which is of growing interest to the data users. Extracting such knowledge is a challenge for information scientists. The area concerned with this task is called data mining [Hand et al., 2001].

Data mining can be defined as 'the analysis of (often large) observational data sets to find unsuspected relationships and to summarise the data in novel ways that are both understandable and useful to the data owner' [Hand et al., 2001].

The datasets examined in data mining are often large. In the case of small datasets, we would rather discuss classical exploratory data analysis. Related issues concern ways of storing and accessing the data, as well as how to determine the representativeness of the data, analyse the data in a reasonable period of time, and decide whether or not an apparent relationship is merely a chance occurrence not reflecting any underlying reality. Data mining is applied to observational data, as opposed to experimental data, that have been collected for some other purposes than the data mining analysis. The strategy used for the data collection therefore does not play any role in the data mining process, which is significantly different from statistical analysis. The patterns found within the data must, naturally, be novel relative to the user's prior knowledge, as there is little point in repeating well-known statements. Another important property of the relationships we seek is simplicity, in order for

them to be comprehensible to the user. Usefulness means that the findings can be made use of or investigated further [Hand et al., 2001].

In a broader context, data mining is a part of knowledge discovery in databases (KDD) [Fayyad et al., 1996; Hand et al., 2001]. This term originates in the field of artificial intelligence. KDD is an iterative and interactive process, which consists of several steps, such as data selection, pre-processing, data enrichment, reduction and projection, data mining, and the interpretation and evaluation of the results of the data mining [Miller and Han, 2001].

Data selection consists of determining a subset in a database for knowledge discovery. Data pre-processing involves noise removal, the elimination of duplicate records, and dealing with the problem of missing data and domain violation. Data enrichment refers to the combination of selected data with other external data. This step may also be included in the pre-processing stage. Data reduction and projection concerns the reduction of both dimensionality and numerosity to get more efficient representations of the information space [Miller and Han, 2001].

Data mining is an essential step in KDD. It involves the application of low-level algorithms for revealing hidden information in a database [Miller and Han, 2001]. Data mining is a complex process that requires high-level human intelligence. The application of automated computational algorithms is insufficient in some cases. Visual methods which use the power of the human eye-brain to detect patterns, structures, and anomalies therefore have a special place in data exploration [Hand et al., 2001].

The final steps refer to evaluating, understanding, and communicating the information discovered in the data mining stage. The boundaries between the steps are not easy to state. The steps may not necessarily be performed in a linear order, and some stages may be skipped or revisited [Miller and Han, 2001].

The entire data mining process, as adapted from [Shekhar and Chawla, 2003], is shown in Figure 2.1. Typically, a domain expert and a data mining analyst are involved in defining and solving a specific problem. The parties agree upon a problem statement in an iterative process. The domain expert provides a clue about the problem he needs to solve and access to the database, while the data mining analyst offers techniques and suitable algorithms to address the problem. The data mining alone is usually a time-consuming process involving the transformation of the database into an algorithm-compatible format. The selection of a technique and the choice of an appropriate algorithm are therefore also a non-deterministic and iterative process.

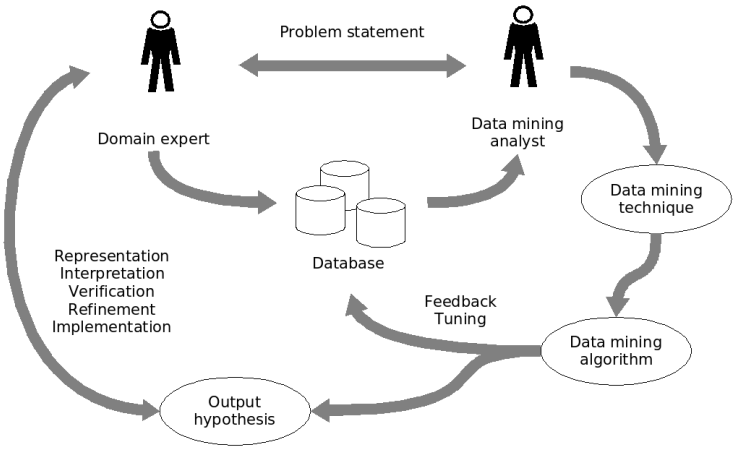



Figure 2.1: Data mining process, adapted from [Shekhar and Chawla, 2003].

The output of the algorithm is typically a potential pattern, which is used to state a hypothesis. The hypothesis needs to be interpreted, verified, and refined together with the domain expert before any decision based on the results of the data mining can be made.

In general, data mining is regarded as a complex concept integrating several research fields, such as artificial intelligence, machine learning, database systems, statistics, or visualisation. Some authors try to keep a distinction between data mining and related techniques. The main differences discussed include the size of datasets, dimensionality, the purpose of the data collection, and issues dealing with uncertainty. In summary, while data mining does overlap considerably with the standard exploratory data analysis techniques of statistics, it also runs into new problems, many of which are the consequences of the size and non-traditional nature of the datasets involved.

Data mining can be described as data-driven hypothesis generation. In this context, it is similar to exploratory data analysis, where the search in the data structures indicates hypotheses about the relationships between cases or variables. However, data mining assumes that novel interesting patterns in very large databases are deeply hidden, so that traditional database queries and statistical methods do not reveal any implicit information [Miller and Han, 2001]. The main difference is in the size of the dataset being analysed, concerning the number of records as well as their

dimensionality. This leads to difficulties in manipulating the data, computational complexity, sampling, and also fundamental problems of restrictions on the choice of models [Hand et al., 2001].

Data mining is typically a secondary process of data analysis, as the data were originally collected for some other purpose. Statistical applications, in contrast, usually deal with primary analysis. The problem with data originally collected to address different problems than those for which they are used is that they may not be ideally suited to these problems. We may also expect additional distortions to occur in the data, such as missing values, contamination, or corrupted data points [Hand et al., 2001]. Statistical models also usually require strict assumptions, such as independence, stationarity, or normality. Data mining, in contrast, goes beyond this traditional domain of statistics [Miller and Han, 2001]. Nonetheless, statistics plays a very important role in data mining, where it is a necessary component [Hand et al., 2001].

One way to view data mining is as a filter step before the application of rigorous statistical tools [Shekhar and Chawla, 2003]. Data mining is more inductive compared to traditional statistical analysis, which is confirmatory. If the information being sought is difficult to specify a priori, then data mining is more appropriate than statistics. However, once a pattern is discovered, data mining cannot compete with statistics in terms of confirmatory power [Miller and Han, 2001].

In general, data mining can be regarded as an interdisciplinary exercise. Statistics, database technology, machine learning, pattern recognition, artificial intelligence, and visualisation all play a role. And just as it is difficult to define sharp boundaries between these disciplines, so it is difficult to define sharp boundaries between each of them and data mining. At the boundaries, one person's data mining is another's statistics, database, or machine learning problem [Hand et al., 2001].

Data mining, as a new technology allowing a wide scope of commercial as well as scientific problems to be tackled, has recently attracted a lot of attention. However, one should not expect it to provide answers to all kinds of questions. Data mining tools provide the potential to lead to valuable results, but, like all discovery processes, with an element of serendipity [Hand et al., 2001].

## 2.2 Mining of geospatial data

Because of its inductive nature and ability to handle large heterogeneous datasets, data mining is an appropriate tool for exploring geographical databases. The requirements for mining geospatial data, however, differ from those for traditional relational databases. The differences result from the special characteristics of geospatial data, such as its geographic measurement framework, spatial autocorrelation, heterogeneity, the complexity of spatial objects and relationships, and the diversity of data types [Miller and Han, 2001].

First, a location in space (and time) must be considered as a single concept, although it consists of up to four dimensions. Representations of geospatial information therefore require the adoption of a topological and geometric measurement framework. While the geometric characteristics of the data relate to the actual location of the object in space [Kraak and Ormeling, 2003], topology covers the properties that are invariant under transformations [Worboys and Duckham, 2004], which represent the spatial relationships among objects [Helokunnas, 1995]. Both geometry and topology affect the attribute values, and are closely related to the definition of distance. The most common framework for distance is Euclidean space; however, there are applications in geographic information science where, for example, travel-time space is more relevant, and transformations of data take place [Miller and Han, 2001].

The nature of geospatial data can be neatly described by Tobler's first law of geography: all objects are related to each other, but closer objects are more related than distant ones. This property is called spatial dependency or autocorrelation. As a consequence of this, the standard assumptions of independence and identically distributed random variables which characterise classical data mining are not applicable for the mining of spatial data [Shekhar and Chawla, 2003].

Geographic attributes often exhibit spatial heterogeneity, meaning the non-stationarity of geographical processes, so that global parameters do not describe the phenomenon well at any particular location [Miller and Han, 2001; Fotheringham et al., 2002].

The complexity of spatio-temporal objects and rules is another distinctive feature. We have to deal with the size, shape, and boundaries of objects, and the relations between them, such as distance, direction, or connectivity. When changing the viewpoint to some different level of detail, we need appropriate scalable tools, which is a

major research topic, especially in the context of mining spatio-temporal data [Miller and Han, 2001].

In geographical data, diverse data types present a unique challenge, as, in addition to the main raster and vector structures, geographic databases increasingly include imagery or geo-referenced multimedia [Miller and Han, 2001].

Because of these special properties, classical data mining techniques have to be modified before being applied to geospatial data. In general, there are two main approaches. The methods used in the first approach consist of processing spatial information as a part of a data mining algorithm. The second approach applies classical data mining algorithms to specifically prepared data, where the spatial relationships have been pre-encoded.

The spatially aware algorithms either correct the underlying statistical model to make it more sensitive to the nature of spatial data, or modify the search function to include the spatial term [Shekhar and Chawla, 2003]. Examples of such applications include spatial autoregressive regression as a classification technique guaranteeing spatial dependencies among objects with the help of a contiguity matrix [Chawla et al., 2001], spatial clustering algorithms [Han et al., 2001], or spatial association and co-location rules associating spatial objects according to their spatial attributes [Koperski and Han, 1995; Malerba et al., 2001].

Mining spatial data using classical data mining methods requires the representation of the spatial term as an additional attribute. Examples of such applications include vertical and horizontal view approaches to modelling space [Estivill-Castro and Lee, 2001], database support for spatial data mining [Ester et al., 1997], expressing spatial dependencies through inferred rules [Santos and Amaral, 2004], or the application of self-organising maps to explore large multidimensional datasets to detect relationships between the attributes [Jiang and Harrie, 2004; Demšar, 2007].

In geographical data mining, the focus is on the spatial aspect of the data. The patterns being discovered involve the spatial and spatio-temporal properties of the individual objects or relationships among them, in addition to non-spatial attributes. The traditional data mining tasks and techniques have their analogues in the geographic data domain. They can be classified according to different criteria, e.g. according to the exploration task into predictive, exploratory, and reductive data mining [StatSoft, 2008], or, on the basis of the extent of the mining, into local and global methods [Mannila, 2002]. The most common categories of data mining tech-

niques described in the literature (e.g. [Miller and Han, 2001; Shekhar and Chawla, 2003]) include the following.

- Spatial segmentation can be divided into spatial classification (classification rules based on the properties of an object, as well as spatial dependency on its neighbourhood) and spatial clustering (based on spatial or non-spatial attributes, and proximity in space-time). Spatial outlier detection can be regarded as an inverse problem to spatial clustering.
- In spatial association rules, the spatial predicates are included in a precedent or an antecedent. The predicates may include topological, distance, or directional relations.
- Spatial trend detection involves finding patterns of change with respect to the neighbourhood of some spatial object.
- Geographic characterisation and generalisation refers to the summarising of data on the basis of spatial and non-spatial attributes. Therefore two types of generalisation are recognised: geographic dominant (spatial aggregation followed by attribute induction to determine spatial patterns) and non-geographic dominant (aggregation of spatial units that share the same high-level description of non-spatial attributes).

Visualisation has also been incorporated as a powerful strategy into the process of mining geospatial data. As a part of the growing discipline of geovisualisation, it integrates cartography, GIS, and scientific visualisation into exploratory tasks such as feature identification, feature comparison, and feature interpretation [MacEachren and Kraak, 2001]. It provides a visual display of the results of complicated computational algorithms and it can also be used directly to detect patterns in data [Demšar, 2006]. In this way, geovisualisation enables human knowledge to be involved in the spatial data mining process.

From the viewpoint of system intelligence [Saarinen and Hämäläinen, 2004], the process of mining geospatial data and knowledge discovery can be regarded as a complex system, where the GI scientist plays the key role of the intelligent and sensitive connection between the domain of information science and various application fields. The GI scientist is required, on one hand, to become familiar with the field of the application and, on the other hand, to be aware of his role in the network of interaction,

bringing a new spatial thinking approach to the application [Mäkelä and Virrantaus, 2008; Virrantaus, 2009]. Such a connecting node is crucial for the communication between the parties involved in all stages of the data mining process, from problem statement, through system tuning, to final hypothesis formulation.

Geographical data mining is currently a research topic. Miller and Han provide an overview of interesting issues in mining geospatial data [Miller and Han, 2001]. They mainly discuss the effects of spatial dependency, database support, techniques for both automated and visual mining, and typical applications for extracting patterns from topographic maps, remotely sensed imagery, and mobile trajectories, which are gaining remarkable attention nowadays. Geographical databases are potentially rich sources of information and spatial data mining offers promising technologies. However, it requires a development on the theoretical level, as well as linking available methods to the reality of GIS.

## 2.3 Representation of knowledge discovered

The data mining process aims to discover knowledge hidden in databases. The term 'knowledge' is used widely, but often quite vaguely, as many connotations and meanings are attributed to it in science, as well as in everyday life [Maier, 2004]. Although philosophers have grappled with the question of what knowledge is over the past two millennia, there is still no consensus on the nature of knowledge, except that it is based on perception that can provide a rational justification for it [Jashapara, 2004].

Ackoff introduces a hierarchy of the content of the human mind as relations between five categories [Ackoff, 1989]:

- data represent objects, events, and/or their properties. They are products of observation;
- information is data that have been processed into useful forms. It provides answers to questions such as 'who', 'what', 'where', and 'when';
- knowledge can be regarded as the application of data and information to answer 'how' questions. It can be obtained from experience or from someone who has obtained it from experience;

- understanding is contained in answers to 'why' questions. It facilitates and accelerates the acquisition of knowledge, and
- wisdom is the ability to perceive and evaluate the long-run consequences of behaviour. While information, knowledge, and understanding contribute primarily to efficiency, wisdom provides an assurance of effectiveness.

The first four categories relate to the past; they deal with what has been or what is known. Only the last category, wisdom, deals with the future because it incorporates vision and design.

Bellinger et al. modify Ackoff's hierarchy, as presented in Figure 2.2. The diagram represents the transitions from data to information, to knowledge, and, finally, to wisdom. Understanding is not a separate level of its own, but a key for the transition from each stage to the next. Knowledge thus represents a pattern that generally provides a high level of predictability as to what is described or what will happen next [Bellinger et al., 2004]. On the basis of the hierarchy that is presented, knowledge can also be interpreted as 'actionable information', which allows us to make better decisions than information or data do [Jashapara, 2004].

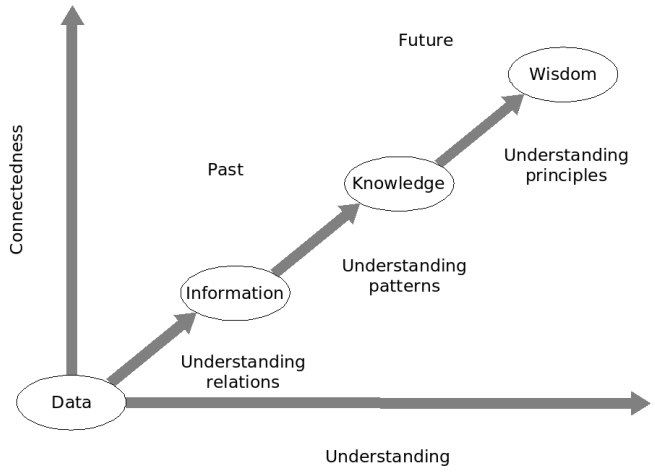



Figure 2.2: Data, information, knowledge, and wisdom hierarchy [Bellinger et al., 2004].

In accordance with the above, knowledge can be broadly defined as a concept

'comprising all cognitive expectancies – observations that have been meaningfully organised, accumulated, and embedded in a context through experience, communication, or inference – that an individual or organisational actor uses to interpret situations and to generate activities, behaviour, and solutions, no matter whether these expectations are rational or used intentionally' [Maier, 2004].

The meaningful representation of observations is crucial for knowledge formation. Different kinds of patterns sought during the data mining process may be characterised in different ways. One such characterisation is the distinction between models, as global summaries of a dataset, and patterns such as statements about restricted regions of space [Hand et al., 2001]. By their nature, local statistics emphasise differences across space, whereas global statistics emphasise similarities across space. Consequently, local statistics are useful in searching for exceptions in the data, which places them in the realm of exploratory data analysis methods, where the emphasis is on developing hypotheses from the data, as opposed to the more traditional confirmatory types of analysis in which the data are used to test a priori hypotheses [Unwin and Unwin, 1998; Fotheringham et al., 2002]. However, in order to detect unusual behaviour we need a description of usual behaviour, so global models and local patterns can sometimes be regarded as opposite sides of the same coin. The distinction is thus not to be a proscriptive constraint [Hand et al., 2001].

According to their approach, observations can be classified as being qualitative or quantitative. While qualitative methods aim to provide a complete description, which is subjective, quantitative methods deal with numbers in an attempt to explain what is observed, allowing an objective comparison between observations [Miles and Huberman, 1994].

We can also draw a distinction between visualising, exploring, and modelling the data. Data visualisation permits us to obtain displays of various descriptions of the data quickly, easily, and flexibly. Exploratory methods involve seeking patterns in the data and thus developing hypotheses about the data by making few a priori assumptions. Explicit statistical models serve to formally test certain hypotheses, or to estimate with some precision the extent and form of the relationships that are of interest. The distinction between visualising, exploring, and modelling is usually not clear-cut, as there is a close interplay between the three, with data being visualised first and interesting aspects being explored, which possibly leads to modelling and further visualisation and exploration to refine the model. The approaches should be

therefore seen as interlinked in an interactive fashion [Bailey and Gatrell, 1995].

Observations can be further grouped according to the exploration framework used during the analysis. Univariate analysis describes the variation within a single variable, bivariate analysis considers the relationship between two variables, and multivariate analysis regards the effects of more variables simultaneously. In the domain of geospatial data, we additionally distinguish spatial, temporal, and spatio-temporal patterns and relationships.

Spatial statistics considers spatial patterns in terms of deviations from random distribution. Various summary statistics describe the first- and second-order properties of point patterns. First-order properties represent variations in the mean value of observed events per unit area in space, suggesting a global or local trend. Second-order properties address the spatial dependence between objects [Diggle, 2003].

Observations organised in a meaningful way are easier to communicate between the parties involved in the data mining process, which supports the understanding of the phenomenon being studied. In this way, data mining enables knowledge to be induced which can be used for decision making or for further scientific investigation.

# Chapter 3

# Description of case application and materials

This study applies selected methods of spatial data mining to analyse the distribution of domestic fires. It demonstrates how each method unveils the relations between the incidents and factors influencing their occurrence for a better understanding of the phenomenon. The study area covers the city of Helsinki.

## 3.1 Domestic fires

Domestic fires, denoting fires in buildings, represent a serious threat and withal a common disaster that the inhabitants of an urban environment may experience. In Finland, fire brigades are called to approximately 3,200 domestic fires every year, which represents almost one third of all fires [Rescue services in Finland, 2006]. Domestic fires cause significant material damage and dozens of them are fatal. Rescue departments recognise the danger and attempt to enhance population safety, focusing on prevention as well as planning the preparedness of fire brigades to minimise the consequences.

Effective emergency planning may benefit from a mature risk model. Currently, the characteristics of domestic fires are not well defined and research is being conducted to improve the understanding of the phenomenon in order to make reliable risk estimates. Tillander, for example, focuses in her work [Tillander, 2004] on quantitative methods for fire risk assessment in buildings in Finland. The work deals with the

ignition frequency and also the consequences of fires under different scenarios. It applies elementary statistical methods to fire records and related building characteristics to identify the main components of fire risks.

The present study, in contrast, focuses on the probability distribution of incidents within the background environment. It applies advanced analytical methods to provide a deep insight into the relations between domestic fires and various underlying aspects, as an understanding of the phenomenon supports realistic estimates of the pattern of the outbreaks of fires.

In constructing a probability model for domestic fires, it is important to identify possible conditioning factors. As most domestic fires are caused by human actions, the spatial distribution of the inhabitants may be an important influence. Besides the population density, other attributes describing the socio-economic structure are also of interest. The occurrence of domestic fires is also related to human activities and is subject to temporal variations over different scales. Other influences may include building characteristics, such as a building type or age. These aspects are analysed to explain the reasons for domestic fires and identify important factors that influence them.

## 3.2 Data description

### 3.2.1 Incident records

The Finnish Fire & Rescue services document all the rescue missions the fire brigades were called to in a Pronto database [Emergency Services College, 2009]. It represents a valuable data source, as it contains a detailed description of each mission, such as the location of the incident in X- and Y-coordinates in addition to the address, time of announcement, incident type, and the response time. Since 2005, the data have been recorded via an electronic report filed by the mission commander and the coordinates of the incident location are inserted by clicking a mouse on the particular place in the map. This process should ensure the highest possible accuracy of the recorded data.

This study focuses on domestic fires reported to Pronto between January 2005 and August 2007. The Helsinki metropolitan area, as the most densely inhabited region in Finland, is of particular interest to the rescue services. The city of Helsinki was therefore selected for the case study. Figure 3.1 shows a dot map of the incidents

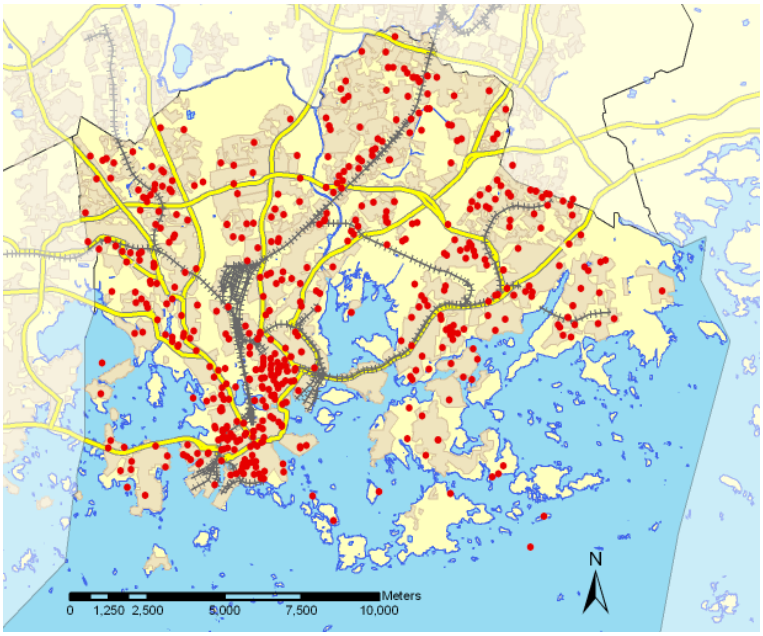


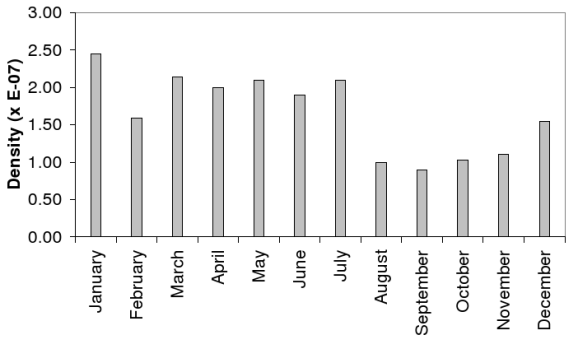

Figure 3.1: Distribution of domestic fires in the study area. The dataset under study consists of 576 records.

studied within the Helsinki area. The dataset that is studied includes 576 records. The locations of the incidents are identified on the basis of X- and Y-coordinates recorded in the database. In addition to their location, all fires are also characterised by their time of occurrence. Figure 3.2 shows changes in the average density (a number of fires per $m^2$) on hourly, day-of-the-week, and monthly scales. As the largest variations occur over the different parts of the day, indicating differences between night fires (n-fires) occurring between 1 a.m. and 8 a.m., day fires (d-fires) occurring between 8 a.m. and 5 p.m. and evening fires (e-fires) occurring between 5 p.m. and 1 a.m., an hourly scale is considered for temporal analysis.

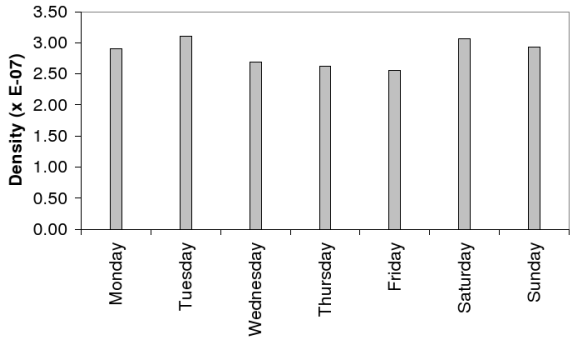
In the case of domestic fires, the Pronto database is designed also to store attributes of the buildings involved. As the records for the data that are analysed are incomplete, an additional dataset is used to provide information about the buildings.

### 3.2.2 Building characteristics

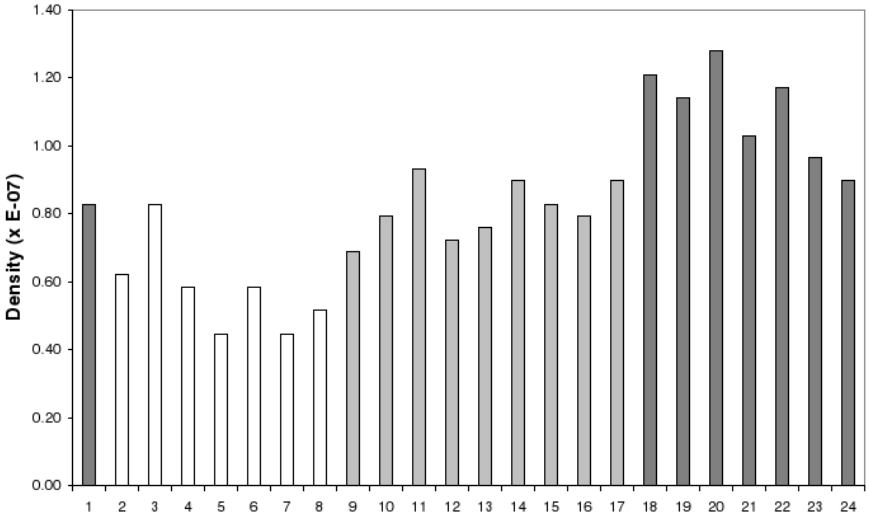
SeutuCD [YTV, 2003], a product of the Helsinki Metropolitan Area Council from 2003, serves as background information. It is a large collection of data on geography,




(a)




(b)




(c)

Figure 3.2: Average density of domestic fires per $m^2$ on different temporal scales: (a) monthly; (b) day-of-the-week; (c) hourly, with colours indicating the d-fires, e-fires, and n-fires categories.

population, and infrastructure that covers the whole metropolitan area, including Espoo, Vantaa, and Kauniainen. SeutuCD provides building data, in particular attributes describing the types of buildings and their ages.

The original dataset contains a detailed description of the buildings based on the purpose of their exploitation. There are 63376 buildings in the study area. The buildings are reclassified into three categories reflecting the main activities of their inhabitants within the city: 'housing', 'leisure', and 'work'.

Buildings have also their year of construction assigned. For those methods of analysis which require the use of categorical data the buildings are classified into three categories: 'old' (built before 1945), 'middle-aged' (constructed between 1945 and 1985), and 'new' (from after 1985). The categories attempt to reflect changes in building construction, as most of the buildings in Helsinki were built after the Second World War, whereas since the late 1980s Finland has been liberalising its economy, which has led to increased construction. The reclassification is described in Table 3.1.

The datasets are associated on the basis of proximity; the fire records get the attributes of the nearest building. Figure 3.3 shows the average density of domestic fires related to particular building type and building age categories within the study area. The values are normalised by the number of buildings falling into particular categories. More fires are related to work buildings than to other types, and more fires occur in buildings from before 1945 than in more recent buildings. These relationships are the subject of further analysis.

### 3.2.3 Census data

Statistics Finland provides census data collected in 2006, which are temporally consistent with the analysed incident records [Statistics Finland, 2006]. They contain detailed information on the socio-economic population profile. Data protection, however, applies to an educational or consumer structure and a labour force if the records relate to fewer than ten people. The data are aggregated in a grid with a cell size 250 $\times$ 250 m. The dataset supplies information about the population and workplace density, unemployment rate, income, level of education, and stage of life in households, reflecting the age of their inhabitants. The study area involves 2496 grid cells.

If the method of analysis applied requires the use of categorical data, census data are classified on the basis of natural breaks, also called Jenks classification [Jenks,

| Building type | Original classification | Nr. of buildings | Percentage |
|---|---|---|---|
| Housing | houses (A) | 44096 | 69.6% |
| | free time houses (B) | | |
| Leisure | travel service buildings (C) | 12823 | 20.2% |
| | reunion buildings (G) | | |
| | stores (K) | | |
| | other buildings (N) | | |
| Work | offices (D) | 6457 | 10.2% |
| | traffic buildings (E) | | |
| | treatment buildings (F) | | |
| | schools (H) | | |
| | industrial buildings (J) | | |
| | fire & rescue buildings (L) | | |
| | agrarian buildings (M) | | |
| Building age | Year of construction | Nr. of buildings | Percentage |
| Old | $< 1945$ | 8427 | 13.3% |
| Middle-aged | 1945-1985 | 32263 | 50.9% |
| New | $\geq 1985$ | 22686 | 35.8% |

Table 3.1: Reclassification of building attributes.

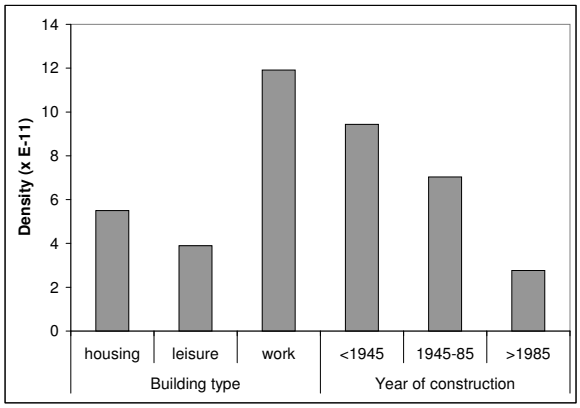


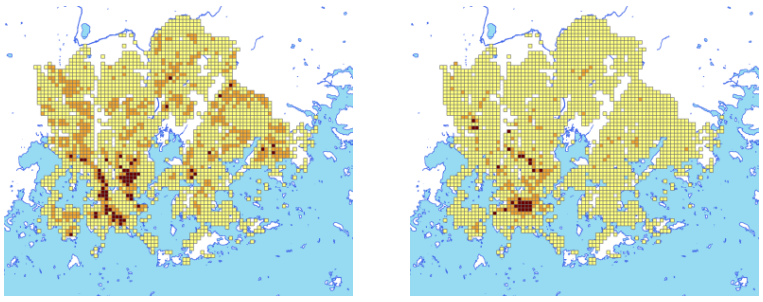

Figure 3.3: Average density of domestic fires per $m^2$ in relation to building attributes. Values are normalised by a number of buildings falling to particular categories.

1967], as described in Table 3.2.

Figure 3.4 presents the distribution of the socio-economic attributes that were studied in the study area. The incident records are associated with the attributes of the census grid cell they fall into. The average density of domestic fires in relation to the socio-economic attributes shown in Figure 3.5 indicates existing relationships that are included in further analysis.

| Population density | Nr. of inhabitants per cell | Nr. of cells | Percentage |
|---|---|---|---|
| Low | $< 304$ | 1876 | 75.2% |
| Middle | 304–960 | 554 | 22.2% |
| High | $\geq 960$ | 66 | 2.6% |
| Workplace density | Nr. of workplaces per cell | Nr. of cells | Percentage |
| Low | $< 530$ | 2325 | 93.1% |
| Middle | 530–1959 | 142 | 5.7% |
| High | $\geq 1959$ | 29 | 1.2% |
| Households with children | Ratio to all households | Nr. of cells | Percentage |
| Low | $< 0.253$ | 1487 | 59.6% |
| Middle | 0.253–0.625 | 922 | 36.9% |
| High | $\geq 0.625$ | 87 | 3.5% |
| Households with adults | Ratio to all households | Nr. of cells | Percentage |
| Low | $< 0.413$ | 956 | 38.3% |
| Middle | 0.413–0.700 | 1146 | 45.9% |
| High | $\geq 0.700$ | 394 | 15.8% |
| Households with pensioners | Ratio to all households | Nr. of cells | Percentage |
| Low | $< 0.171$ | 1416 | 56.7% |
| Middle | 0.171–0.447 | 990 | 39.7% |
| High | $\geq 0.447$ | 90 | 3.6% |
| Incomes | Yearly incomes | Nr. of cells | Percentage |
| Low | $< 43650$ | 840 | 48.5% |
| Middle | 43650–126287 | 838 | 48.4% |
| High | $\geq 126287$ | 53 | 3.1% |
| Unemployment | Unemployment rate | Nr. of cells | Percentage |
| Low | $< 0.078$ | 983 | 57.0% |
| Middle | 0.078–0.179 | 621 | 36.0% |
| High | $\geq 0.179$ | 120 | 7.0% |
| Education | Ratio of basic education | Nr. of cells | Percentage |
| Low | $> 39.7\%$ | 742 | 41.6% |
| Middle | 24.2–39.7% | 723 | 40.6% |
| High | $\leq 24.2\%$ | 317 | 17.8% |

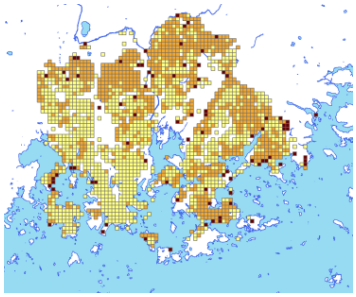
Table 3.2: Classification of census data.



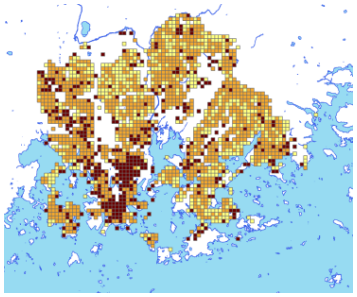
(a) Population density



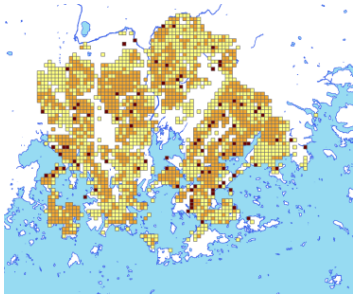
(b) Workplace density



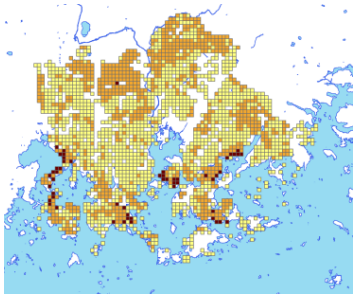
(c) Households with children



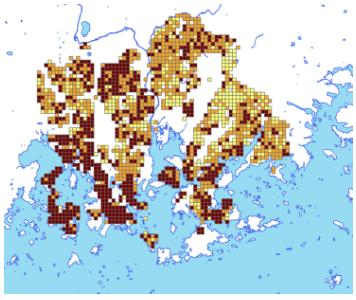
(d) Households with adults



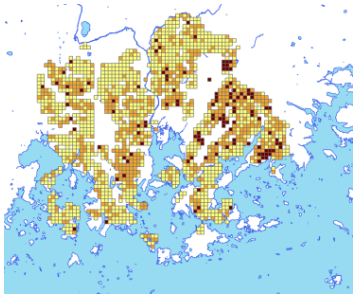
(e) Households with pensioners



(f) Incomes



(g) Education



(h) Unemployment

Figure 3.4: Distribution of socio-economic population attributes in the study area. The colour scale from red to yellow follows the attribute values from high to low. Blank cells represent protected data.

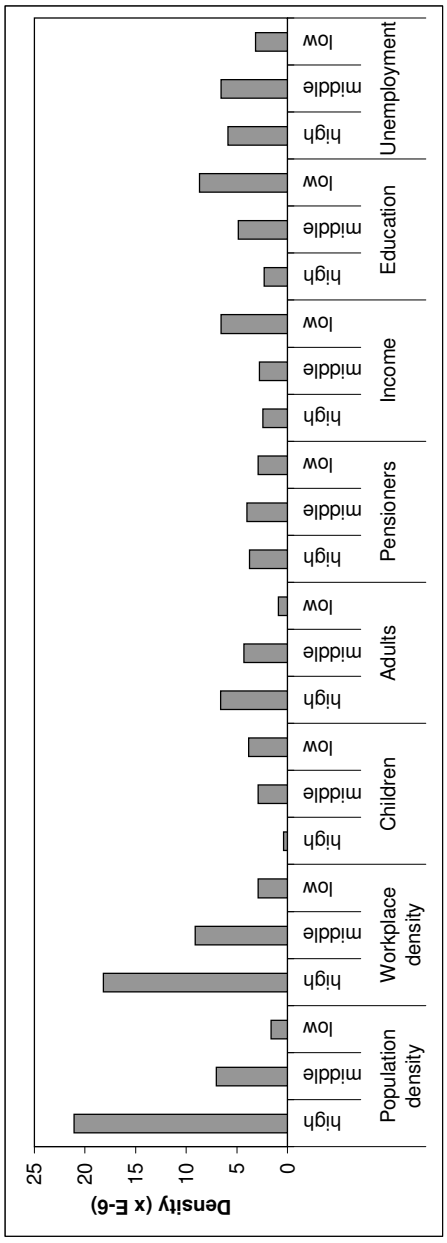



Figure 3.5: Average density of domestic fires per $m^2$ in relation to socio-economic aspects. Values are normalised by a number of census cells falling to particular categories.

# Chapter 4

# Visual data mining

This chapter is based on the study presented in the following article. The author (formerly Křemenová) contributed to the theoretical studies of spatial data mining, the exploration of the data, and the writing of the paper.

Demšar, U., Krisp, J., Křemenová, O. (2006) Exploring geographical data with spatio-visual data mining. Riedl, A., Kainz, W. and Elmes, G. (eds.): Progress in Spatial Data Handling, Proceedings of the 12th International Symposium on Spatial Data Handling, Springer-Verlag, Berlin Heidelberg.

## 4.1 Method

Visualisation represents a graphic communication of information, data, documents, or structures [Fayyad et al., 2002]. Following [DiBiase, 1990; MacEachren and Ganter, 1990], the role of visualisation in the context of geographical data moves from simple presentation of the results to the earlier stages of exploration and confirmation. Information visualisation methods use an interactive visual representation of abstract data to amplify cognition [Schneiderman and Plaisant, 2005]. They take advantage of the ability of the human visual perception system to detect patterns in graphical images. Interactivity is a key factor in visualisation. It forms a dynamic exploration system which allows the user to interact with the data [Grinstein and Ward, 2002]. In this way, information visualisation enables us to provide an insight into data that are large and complex and thus contributes to the process of knowledge discovery.

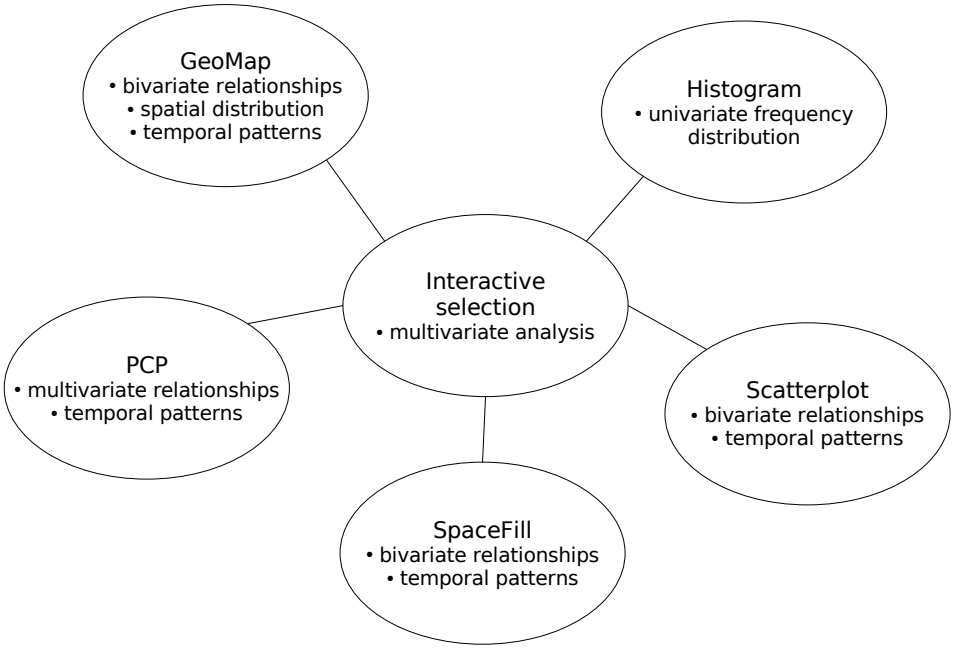



Figure 4.1: Visual data mining system.

In contrast to the complex mathematical background of computational data mining methods, visualisation is also easier for the user to understand [Keim, 2001].

The visual data mining methods used in this chapter form an interactive and interconnective exploratory system, as presented in Figure 4.1. The system consists of a bivariate map, a parallel coordinates plot (PCP), and a multiform bivariate matrix with histograms, scatterplots, and spaceFill visualisations. The visualisations allow interactive data selection and brushing – highlighting of the selected objects in all other visualisations. Such a dynamic iterative exploration provides a better visual impression and easier pattern recognition [Takatsuka and Gahegan, 2002; Gahegan et al., 2002].

A bivariate choropleth map shows the geographical extent of the data. A colour scheme assigned to a map based on two variables is transferred to other visualisations in the system that do not have their own colour schemes [Takatsuka and Gahegan, 2002; Gahegan et al., 2002].

A multiform bivariate matrix [MacEachren et al., 2003] is used for bivariate data analysis. Data dimensions are assigned to the two axes of the bivariate matrix and the matrix elements represent a corresponding bivariate data visualisation. In this case the multiform bivariate matrix consists of scatterplots and spaceFill visualisations above and below the diagonal, respectively. In addition, data histograms representing the data distribution in the form of tabulated frequencies are located on the diagonal.

Scatterplots visualise the data as points in a 2-D display, where the two axes represent the respective attributes being analysed. They are suitable for discovering correlations between the two attributes. However, when there are too many data points the predicate value of the scatterplots is reduced as a result of overprinting [Hand et al., 2001].

As each grid in a spaceFill visualisation represents one data element, it solves the problem of overprinting [MacEachren et al., 2003]. The order of the grids is based on the attribute values of one of the two variables being considered, while their colour is assigned according to the attribute values of the other variable. Several orders are possible for the grids: a scan line is used in this application, starting from the lowest values in the bottom left-hand corner and proceeding along a scan line towards the highest values in the top right-hand corner. Any regularity in the spaceFill pattern thus indicate a correlation between the two variables.

A parallel coordinate plot [Inselberg, 2002; Edsall, 2003] is finally used for multivariate analysis. It maps a multidimensional space onto a two-dimensional display of a set of vertical equidistant axes, which correspond to the attributes of the data. Each data item is represented by a polyline connecting the axes, while intersecting them at appropriate attribute values. A PCP represents an effective way to unveil the characteristics of data distribution, such as clusters or correlations. However, a PCP suffers from overprinting when displaying large amounts of data with similar values.

The system is built using GeoVISTA Studio, a Java-based collection of visual data mining methods for the scientific exploration of geographical data [Takatsuka and Gahegan, 2002; Gahegan et al., 2002].

## 4.2 Data pre-processing

In order to encode the spatial relationships in the architecture of the dataset, the geographical space is represented as grid layers, one for each attribute. The grids

have identical resolutions, with each grid cell symbolising a neighbourhood unit. The cell size for this application is chosen to be 250×250 m, corresponding to the source census data. The grid layers are integrated into an exploration dataset by the vertical view approach to spatial data mining [Koperski and Han, 1995; Estivill-Castro and Lee, 2001]. This approach is based on a map overlay of all the grid layers. In the resulting database of neighbourhoods, each spatial cell corresponds to a traditional record with associated attributes.

To minimise information loss coming from rasterisation [Zhang and Goodchild, 2002] the incidents are represented as a continuous density surface. The Kernel method [Silverman, 1986] is used for the density estimation. A kernel bandwidth of 200 m is selected in this study. In addition, densities for n-fires (1 a.m. – 8 a.m.), d-fires (8 a.m. – 5 p.m.) and e-fires (5 p.m. – 1 a.m.) are constructed to analyse changes in the relationships on the temporal scale. Building attributes are not related to particular incidents, but to a grid cell in this case. The data are generalised to represent the average age of buildings, respectively the most frequent building type within a grid cell. Original continuous data are used without any additional categorisation.

## 4.3 Results

The bivariate map shows the spatial distribution of domestic fires and the background attributes. As the colour scale depends on two attributes, it is useful to analyse bivariate relationships. Figure 4.2 shows the distribution of e-fires in relation to d-fires. These time categories exhibit a different spatial pattern. While the darker colours indicate areas with a high density of both e- and d-fires, there are areas with a high density of e-fires and a low density of d-fires at the same time (red) and vice versa (green). Such an observation supports the intention to analyse the selected temporal categories separately.

The bivariate matrix is useful for observing the correlations between two variables in the form of scatterplots and spaceFill visualisations; see Figure 4.3, where the relationships under study are marked in red. The colour scale of the spaceFills from pink to green indicates low to high census attributes values, while the arrangement of their cells corresponds to the changes in the density of domestic fires. The smoothness of the transition from pink to green from the bottom to the top (or from the

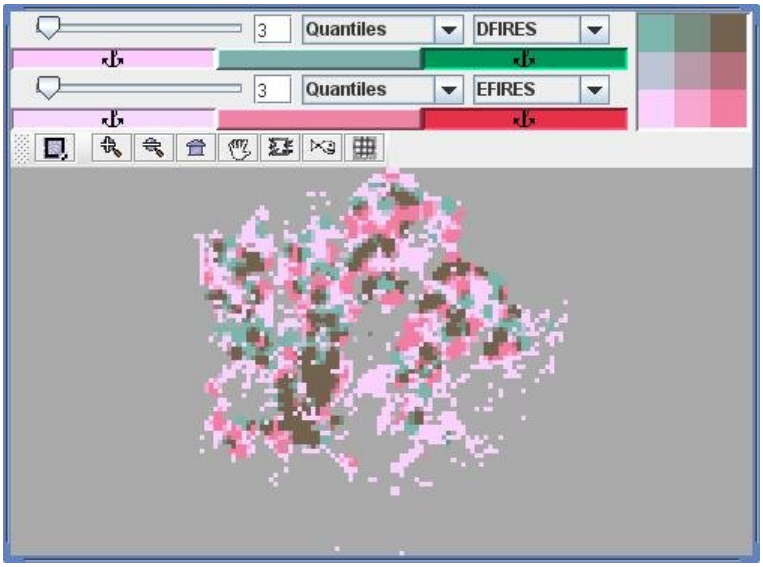
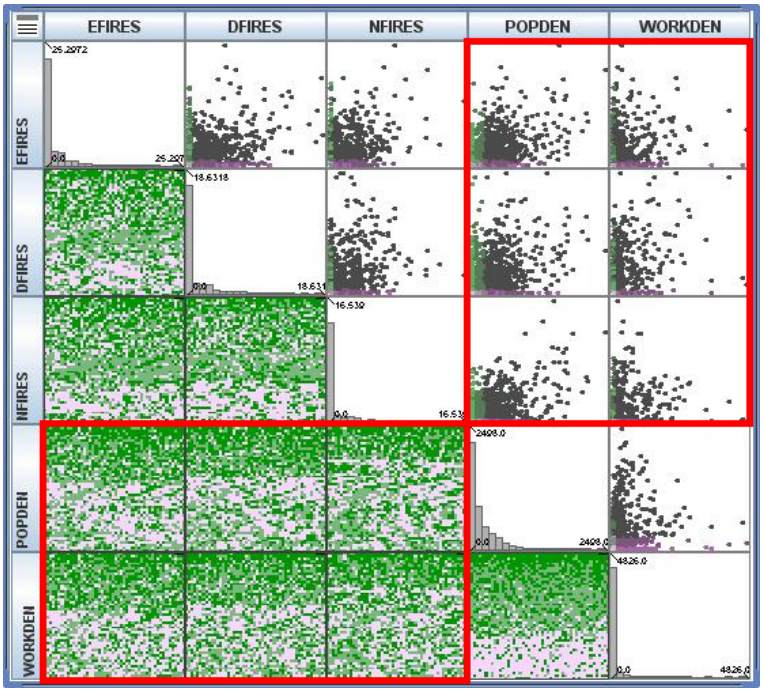



Figure 4.2: Distribution of d-fires and e-fires. Colour scale from pink to green and red indicates low to high density of d-fires and e-fires, respectively. Dark colour indicates high density of both d-fires and e-fires.

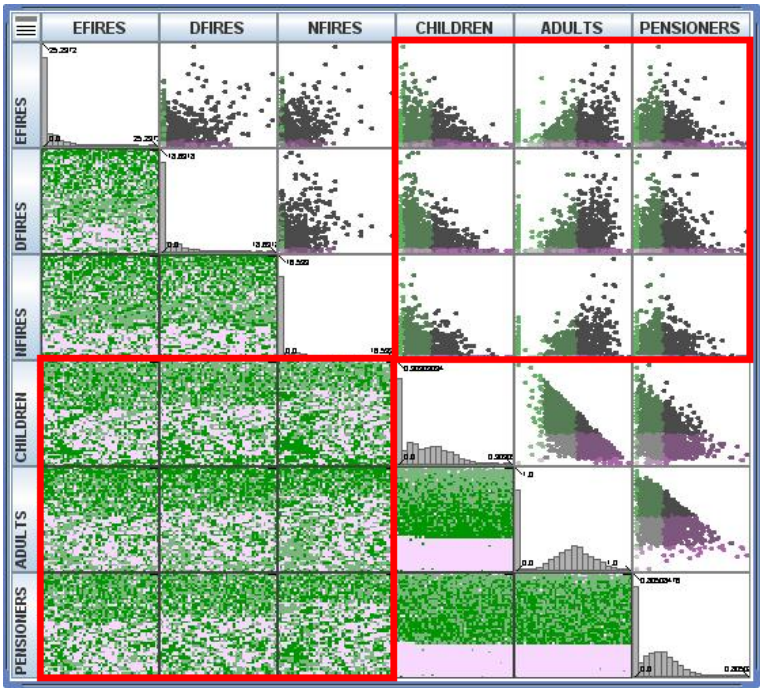
top to the bottom in cases of negative correlation) indicates the smoothness of the correlation. The shape of the patterns in the scatterplots and the arrangement of the colour scale in the spaceFills indicate a positive correlation between the domestic fires and population and workplace density (Figure 4.3 (a)). An additional correlation is observed between domestic fires and attributes describing the structure of households. Whereas households with adults exhibit a strong positive correlation to all temporal categories of fires, the correlation is negative in the case of households with children and weakest in the case of households with pensioners (Figure 4.3 (b)).

Figure 4.4 illustrates the spatial distribution of domestic fires in relation to population density. We can identify areas with a high population density and a low density of fires at the same time (red) and, conversely, areas with a low population density and a high density of fires (green areas). We can also observe temporal changes in the distribution of fires: compared to the daytime, fires move with a high density to the north-western part of the study area in the evening. The density of fires at night is lower.

The techniques discussed above consider the relationships in a purely bivariate

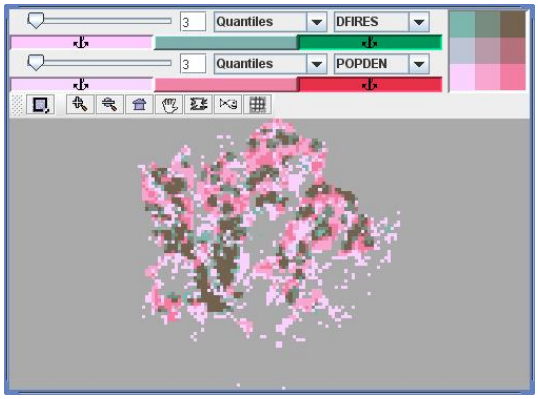



(a)

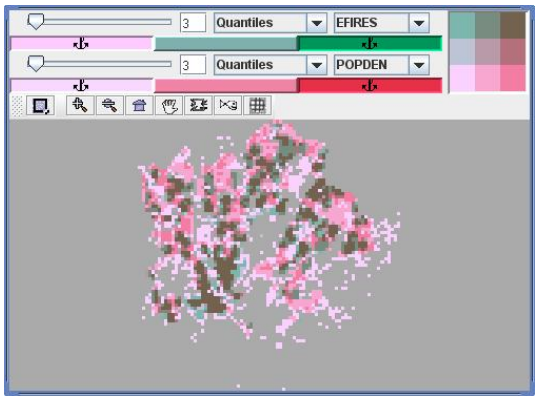



(b)

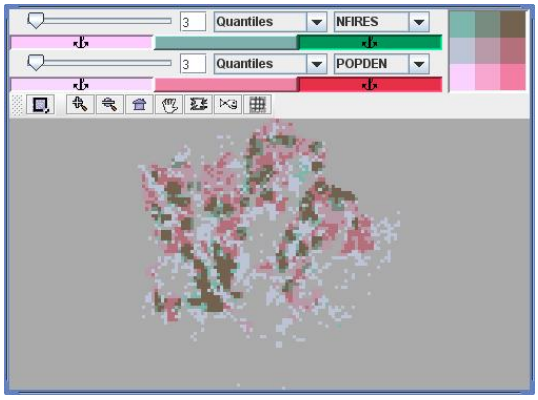
Figure 4.3: Bivariate matrix for the temporal fire categories: (a) population and workplace density; (b) household structure. The colour scale of the spaceFills from pink to green indicates low to high census attributes values, while the arrangement of their cells corresponds to the changes in the density of domestic fires.




(a)




(b)




(c)

Figure 4.4: Temporal changes in the spatial distribution of domestic fires in relation to population density: (a) d-fires; (b) e-fires; (c) n-fires. Colour scale from pink to green and red indicates low to high density of domestic fires and background attributes, respectively. Dark colour indicates a high density of both domestic fires and background attributes.

manner, which can often be misleading. As the PCP enables us to observe all the attributes in detail at the same time, it provides a deeper insight into the data. As the positioning of the axes may lead to different interpretations, an interactive and iterative adjustment of the PCP is necessary. Figure 4.5 shows the respective attribute values for selected high densities of d-fires (a), e-fires (b), and n-fires (c). The colour scale from green through yellow to red in the PCP is assigned on the basis of the respective fire category and high density values are highlighted. In all cases, a high density of fires relates to a high ratio of households with adults, but a low ratio of households with children and a low to medium ratio of households with pensioners. Fires occur mostly in areas where the inhabitants have low incomes. A higher density of workplaces seems to be relevant to d-fires. Similarly, building types classified as work buildings (Type 3) are highlighted in the case of d-fires. In addition, all fire categories correlate with housing buildings (Type 1), in contrast to leisure ones (Type 2).

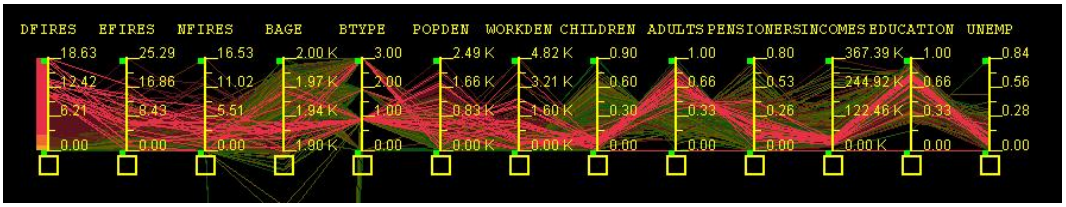
The PCP also enables us to zoom in on attribute values of interest, as illustrated in Figure 4.6. We are interested in buildings built after 1900. The colour scale of the PCP is assigned according to the age of the building. The green colour on the axes representing the density of fires illustrates the influence of the age of the building on an increase in the fire density.

Using interactive selection in several visualisations allows us to observe multivariate relationships. As an example, the highest ratio of households with adults is selected in the PCP. This selection is also displayed in the map; see Figure 4.7. The green-yellow-red colour scale used in both the PCP and the map represents the density of fires. We can observe that areas with a high density of adult households are clustered in the central part of the study area and suffer from high density of domestic fires.

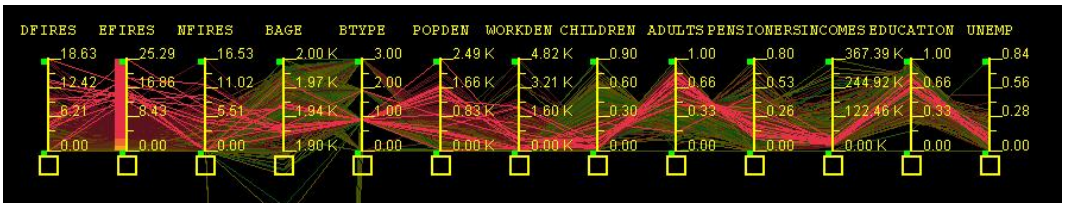
## 4.4 Conclusions

### 4.4.1 Capturing spatial and temporal aspects

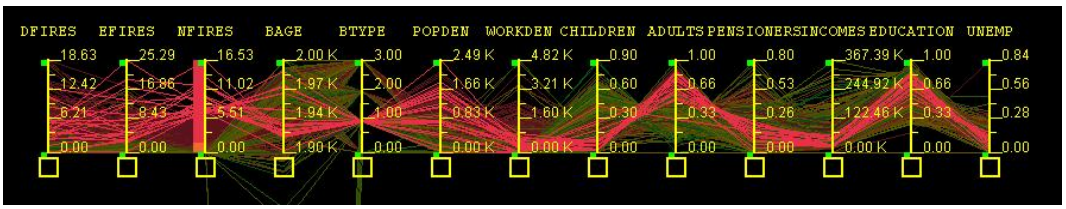
Visual methods for data mining are valuable for detecting patterns in large and complex data. Provided there is suitable conceptualisation of spatial relationships, they can be effectively applied to geographical data. In this study, the neighbourhood




(a)




(b)




(c)

Figure 4.5: PCP for high density of domestic fires: (a) d-fires; (b) e-fires; (c) n-fires. The colour scale from green through yellow to red is assigned on the basis of the respective fire category and high density values are highlighted.

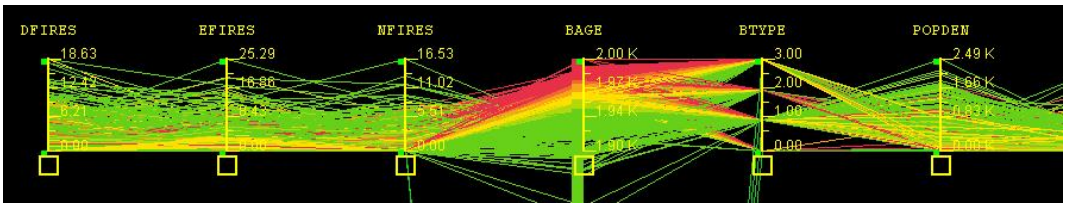
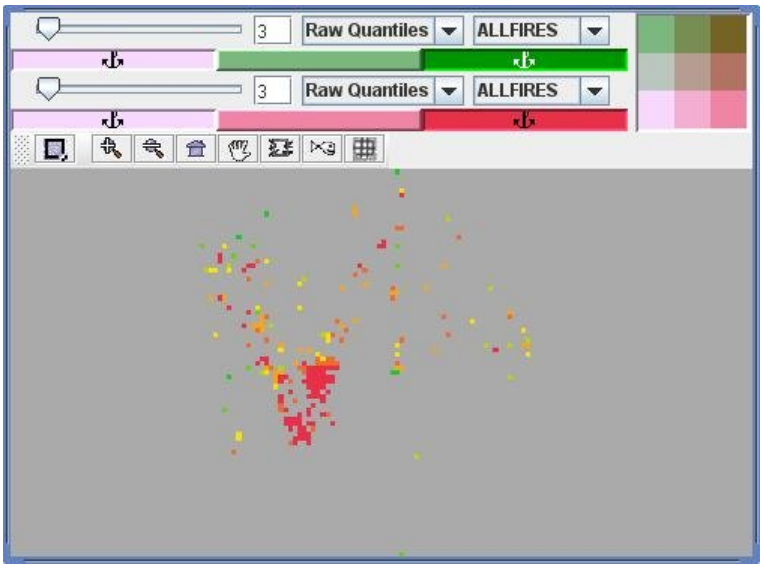
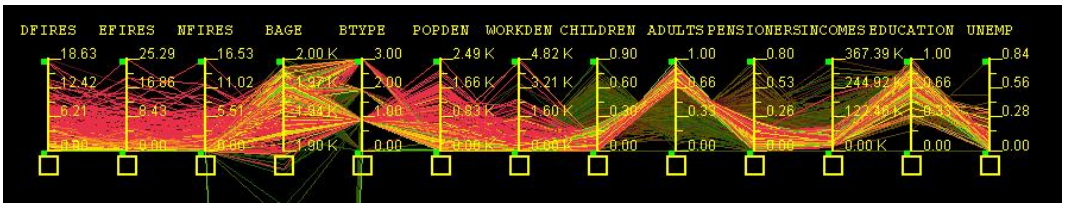



Figure 4.6: PCP for density of domestic fires and building age. The colour scale assigned according to the age of the building illustrates the influence of the age of the building on an increase in the fire density.




(a)




(b)

Figure 4.7: Connection between the PCP and the map. The highest ratio of households with adults is selected in the PCP (b) and the selection is also displayed in the geoMap (a). The colour scale from green through yellow to red indicates low to high density of all fires.

relationships are represented by a grid cell, but other options, e.g. buffer zones, are possible. The use of continuous density surfaces instead of raw cell counts partially overcomes the problem of rasterisation. An appropriate resolution for the analysis must be chosen with respect to the application.

The methods are also suitable for visualising temporal patterns. In this study, the temporal analysis is performed on a pre-defined scale, selected according to prior knowledge about the temporal distribution of the attributes being studied.

When exploring spatial data, a connection to the original geographical space is useful. The exploratory system used in this study contains a map as one of the visualisations, which helps to identify spatial patterns. Another important feature of the system is interactive selection and brushing, which makes the effective exploration of the dataset possible.

### 4.4.2 Type of knowledge discovered

The methods for data visualisation presented here allow univariate, bivariate, and multivariate patterns in the data to be discovered. With suitable spatial pre-processing, they can also be used to unveil spatial and spatio-temporal patterns. The bivariate choropleth map shows the spatial distribution of the variables studied, while two attributes can be studied at the same time by applying a suitable colour code. In addition, it enables the spatial distribution of the patterns selected in other visualisations to be observed. Both scatterplots and spaceFills enable bivariate relationships to be detected. The former may be more intuitive to the user, but the impression suffers from overprinting. The latter visualisation avoids overprinting and thus provides an estimate of the strength of the correlation. The PCP is useful for an insight into the multivariate relationships, while allowing particular attribute values to be focused on. The detailed multivariate analysis is based on interactive selection and visualisation of the patterns in several visualisations.

In general, visual methods do not quantify the observed relationships; they rather serve for the formulation of hypotheses. However, the strength of correlations can, to some extent, be visually estimated. The main advantage of visual methods lies in providing a fast insight into the data.

### 4.4.3 Weaknesses

The visual approach to the data mining provides a qualitative overview of the data. It enables patterns and relationships in the data to be detected. The significance of the patterns and the strength of the relationships can only be estimated visually. Further analysis is necessary to quantify the findings.

The bivariate matrix and map alone analyse the relationships in a purely bivariate manner, which reveals the association or differences between two variables. Although knowledge about bivariate relationships may be useful for the user, the phenomenon being studied is expected to be more complex and such an approach to the analysis can be very misleading. In order to understand the message the results bring, the user should be aware of the limitations the bivariate analysis involves. The analysis should also continue by exploration of the multivariate relationships.

Information visualisation provides a wide range of methods for data mining, from simple and intuitive scatterplots to more advanced ones, such as spaceFills or a PCP, the effective use of which generally requires training. The more detailed the insight into the processes underlying the data that is required, the more demanding the interpretation is. The PCP enables the processes to be explored in detail; however, it may be difficult to use. The positioning of the axes, and also their scales, can lead to different interpretations. An interactive and iterative adjustment of the PCP is necessary before the user can unveil useful patterns. A detailed multivariate analysis is also supported by the principle of interactive selection and brushing, when the selected patterns are highlighted in several other visualisations. A careful exploration performed by an experienced user is required in order to describe the processes underlying the data.

### 4.4.4 Requirements for the user

Capturing spatial aspects with conventional exploratory methods always requires some degree of pre-processing. As this step is application-dependent, it cannot be automated. A GIS expert is therefore necessary to perform the pre-processing, with regard to both the conceptual and technical levels.

Information visualisation uses sophisticated methods, which may be confusing for an inexperienced user. Training is usually necessary before their potential can be fully exploited. On the other hand, visualisation does not require an understanding

of difficult algorithms behind them. The results are therefore easy to present to the public.

Geographical visualisation is supported by an open software development environment, GeoVISTA Studio [Takatsuka and Gahegan, 2002; Gahegan et al., 2002], which allows a dynamic analysis and display of geographical data within a modularly designed interface.

# Chapter 5

# Contingency tables

The following article relates to the method presented in this chapter. The author contributed to the theoretical studies, designed the conceptual framework of the analysis, performed the analysis and also played a leading role in writing the paper.

Špatenková, O. & Krisp, J. (2007) The use of contingency tables to value variables for spatial models. Proceedings of the 5th International Symposium on Spatial Data Quality, Enschede, the Netherlands, 2007.

## 5.1 Method

Contingency tables are broadly used for the statistical analysis of categorical data [Bishop et al. 1975; Everitt, 1992; Agresti, 2002]. They represent the frequencies of occurrence of particular combinations of two or more discrete variables, although two-way contingency tables considering only two variables at a time are usually the matter of interest. The term 'contingency', as introduced by Karl Pearson, refers to the associations between the variables [Pearson, 1904]. A hypothesis testing whether the variables are independent or whether there is an association among them is frequently an issue addressed by contingency tables. A tabular representation of the data distribution also allows the biggest disproportions to be identified and a deeper insight to be gained into the relationships in the data.

Let $X$ and $Y$ be two discrete variables that are exclusive (categories do not overlap) and exhaustive (categories include all possibilities) and that can be assigned a

finite number of values $1, \ldots, r$ and $1, \ldots, c$ respectively. The contingency table $(n_{ij})$ represents the empirical frequencies of a particular combination of the variable values in a sample dataset. It can be expressed in the form of a matrix, as shown in Table 5.1. The corresponding column sums $n_{.j}$ and row sums $n_{i.}$ are called marginal frequencies or marginal totals. The sample size $n$ is also called the grand total.

| $X/Y$ | 1 | ... | $c$ | $\sum$ |
|---|---|---|---|---|
| 1 | $n_{11}$ | ... | $n_{1c}$ | $n_{1.}$ |
| ... | ... | ... | ... | ... |
| $r$ | $n_{r1}$ | ... | $n_{rc}$ | $n_{r.}$ |
| $\sum$ | $n_{.1}$ | ... | $n_{.c}$ | $n$ |

Table 5.1: A contingency table.

Testing the independence of the two variables is one of the most frequent analysis tasks. The two variables are independent if and only if the probability distribution $p_{ij} = p_{i.}p_{.j}$ is true for all the couples $i = 1, \ldots, r$ and $j = 1, \ldots, c$, in other words if the conditional probability is equal to the unconditional probability. The probability $p_{ij}$ can be approximated by the relative observed frequencies $\frac{n_{ij}}{n}$. Thus, if the two variables being studied are independent, the relationship $\frac{n_{ij}}{n} \sim \frac{n_{i.}}{n}\frac{n_{.j}}{n}$ is valid. The question is to what extent variations existing within an empirical sample can be considered approximately equal. The answer is given by Pearson's chi-square test. The variable

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{5.1}$$

has a $\chi^2$ distribution with $(r-1)(c-1)$ degrees of freedom. Thus, when we get $\chi_0^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$, where $\alpha$ is the level of significance, we reject the null hypothesis about the independence of the two variables.

Note that the $\chi^2$ test may not be reliable if some cell expectations are too small, as the corresponding term in $\chi_0^2$ tends to dominate over the others because of its small denominator. The most conservative recommendation says that all expected cell counts should be 5 or more; however, it is considered to be too restrictive. Cochran's standard rule allows the $\chi^2$ test to be used for contingency tables that have at least 80% of the expected cell counts 5 or more, while the minimum expected cell count is 1 [Cochran, 1954].

The value of $\chi_0^2$ measure serves only as a basis for testing the hypothesis of independence; however, it does not correspond to the degree of association. To compare the associations between different contingency tables, various normings of $\chi_0^2$ measure have been proposed. The Cramér's $V$ coefficient:

$$V = \sqrt{\frac{\chi_0^2}{n \quad \{min[(r-1),(c-1)]\}}} \tag{5.2}$$

is suitable, as it can be applied to general (asymmetric) contingency tables. Its values lie between 0 and 1, while the maximum is reached for a complete association.

Besides the identification of associations in the data and assessing the strength of the relationships, the analysis of contingency tables can reveal more details describing the data. For example, we can also pinpoint the table cells invoking rejection of the independence hypothesis. Such findings provide a deeper insight into the data and enable a better understanding of the existing relationships to be gained.

The analysis of contingency tables is supported by common statistical software nowadays. Within this study, the capability of MS Excel was used as it is a widespread program familiar to common users.

## 5.2 Data pre-processing

The analysis of contingency tables is a traditional statistical method. In order to apply it to spatial relationships, the geographical space being studied is represented by grid layers, as in the previous chapter. The layers are connected on the basis of a map overlay so that each spatial cell corresponds to a traditional record with its associated attributes [Koperski and Han, 1995; Estivill-Castro and Lee, 2001]. Zero valued cells are excluded from the analysis. The cell size is chosen to be 250×250 m, which corresponds to the source census data. Additional building attributes are related to particular grid cells. They represent the average age of buildings and the major building type within a grid cell.

The incidents are represented as a continuous kernel density surface [Silverman, 1986]. A kernel bandwidth of 200 m is selected in this study. As the analysis of contingency tables requires categorical variables, the incident density is classified into three classes representing areas with high, middle, and low densities of domestic fires according to Jenks' classification [Jenks, 1967].

For a temporal analysis, the incident dataset is split according to density variations into three categories representing n-fires (1 a.m. – 8 a.m.), d-fires (8 a.m. – 5 p.m.), and e-fires (5 p.m. – 1 a.m.). These categories are analysed individually.

## 5.3 Results

The contingency tables for domestic fires and the attributes that are studied are constructed for the entire dataset and also for the three temporal categories. The null hypothesis of independence is tested according to the $\chi^2$ test (Table 5.2). As the size of the tables is $3 \times 3$, which corresponds to 4 degrees of freedom, the $\chi^2$ values are compared to the $\chi^2_4(0.05) = 9.49$ for a test at a significance level of 5%.

| | all fires | n-fires | d-fires | e-fires |
|---|---|---|---|---|
| population density | 521.37 | 393.92 | 287.08 | 325.41 |
| workplace density | 345.46 | 103.84 | 224.13 | 172.52 |
| children | 151.24 | 60.74 | 108.98 | 74.27 |
| adults | 214.50 | 100.90 | 119.54 | 101.96 |
| pensioners | 7.59 | 6.25 | 12.95 | 8.41 |
| incomes | 148.68 | 39.45 | 87.98 | 68.53 |
| education | 40.80 | 7.91 | 5.47 | 16.10 |
| unemployment | 94.49 | 22.16 | 53.15 | 52.84 |
| building type | 18.28 | 22.50 | 29.13 | 19.01 |
| building age | 245.43 | 95.17 | 188.06 | 60.42 |

Table 5.2: Results of the $\chi^2$ test.

The null hypothesis of independence is rejected for most of the variables studied, which means that the relations observed between domestic fires and the influences studied are statistically significant. Households with pensioners represent an exception, as a correlation is observed only in relation to d-fires. The education of the inhabitants does not correlate with d- and n-fires, either.

The Cramér's $V$ coefficient is calculated to compare the strength of the association between particular variables (table 5.3), while higher values indicate stronger correlations. The Cramér's $V$ values therefore suggest the importance of the variables that are studied to the occurrence of domestic fires. In addition, a comparison of the

Cramér's $V$ coefficients between the n-, d-, and e-fires indicates the variation in the influence of particular variables during the day.

Population density is identified as the main factor influencing the occurrence of domestic fires for all temporal categories, while the correlation is strongest at nights. The density of workplaces is the next most important influence. The correlation between the density of workplaces and fires is most significant during the daytime and least at night. Other significant influences include households with adults, the age of the building, and the incomes of its inhabitants.

| | all fires | n-fires | d-fires | e-fires |
|---|---|---|---|---|
| population density | 0.356 | 0.309 | 0.264 | 0.281 |
| workplace density | 0.289 | 0.158 | 0.243 | 0.204 |
| children | 0.192 | 0.134 | 0.163 | 0.122 |
| adults | 0.229 | 0.157 | 0.171 | 0.158 |
| pensioners | 0.043 | 0.039 | 0.056 | 0.045 |
| incomes | 0.207 | 0.107 | 0.159 | 0.141 |
| education | 0.107 | 0.047 | 0.039 | 0.067 |
| unemployment | 0.166 | 0.080 | 0.124 | 0.124 |
| building type | 0.061 | 0.068 | 0.077 | 0.062 |
| building age | 0.222 | 0.138 | 0.194 | 0.110 |

Table 5.3: Cramér's $V$ values.

A further study of the contingency tables provides a more detailed insight into the relations, as, for example, particular cells invoking high values of $\chi^2$ in the test can be identified. In addition, a comparison of the observed and expected values indicates if the correlation is positive or negative.

Tables 5.4-5.6 demonstrate the contingency tables and calculation of the $\chi^2$ test for the most significant influences, i.e. population density, the density of workplaces, and the density of households with adults for the full dataset as well as for the three temporal subsets. The highest number in the $\chi^2$ test in Table 5.3(b) indicates that there is a strong correlation between a high density of domestic fires and a high population density for all the temporal categories studied. In all cases, the observed pattern is more frequent than the expected values, which implies positive correlations.

A high density of domestic fires also correlates to a high density of workplaces (see

Table 5.4(b)). However, differences are observed during the day, as the occurrence of n-fires relates to a middle density of workplaces. The correlations are positive in this case too.

In the case of households with adults (Table 5.5(b)), the strongest positive correlation to domestic fires is observed between the highest density categories for all temporal subsets. In addition, a low density of adult households is also strongly correlated to high and middle densities of domestic fires, but the correlation is negative. It means that in areas with low and middle densities of households with adults, there is a significantly lower occurrence of domestic fires than would be expected from random distribution.

(a) contingency table

| | | population density | | | |
|---|---|---|---|---|---|
| | | high | middle | low | $\sum$ |
| all fires density | high | 30 | 39 | 17 | 86 |
| | middle | 29 | 251 | 267 | 547 |
| | low | 7 | 264 | 1154 | 1425 |
| | $\sum$ | 66 | 554 | 1438 | 2058 |
| n-fires density | high | 29 | 32 | 16 | 77 |
| | middle | 16 | 132 | 168 | 316 |
| | low | 21 | 390 | 1254 | 1665 |
| | $\sum$ | 66 | 554 | 1438 | 2058 |
| d-fires density | high | 25 | 57 | 35 | 117 |
| | middle | 27 | 184 | 239 | 450 |
| | low | 14 | 313 | 1164 | 1491 |
| | $\sum$ | 66 | 554 | 1438 | 2058 |
| e-fires density | high | 28 | 59 | 32 | 119 |
| | middle | 28 | 205 | 290 | 523 |
| | low | 10 | 290 | 1116 | 1416 |
| | $\sum$ | 66 | 554 | 1438 | 2058 |

(b) $\chi^2$ test

| | | population density | | | |
|---|---|---|---|---|---|
| | | high | middle | low | $\sum$ |
| all fires density | high | 269.08 | 10.85 | 30.90 | 310.83 |
| | middle | 7.48 | 73.10 | 34.73 | 115.31 |
| | low | 32.77 | 37.29 | 25.17 | 95.23 |
| | $\sum$ | 309.33 | 121.24 | 90.80 | 521.37 |
| n-fires density | high | 285.04 | 6.13 | 26.56 | 317.73 |
| | middle | 3.40 | 25.90 | 12.63 | 41.92 |
| | low | 19.66 | 7.56 | 7.06 | 34.27 |
| | $\sum$ | 308.09 | 39.59 | 46.24 | 393.92 |
| d-fires density | high | 120.32 | 20.65 | 26.74 | 167.71 |
| | middle | 10.95 | 32.62 | 18.10 | 61.66 |
| | low | 23.92 | 19.46 | 14.33 | 57.70 |
| | $\sum$ | 155.18 | 72.73 | 59.16 | 287.08 |
| e-fires density | high | 153.25 | 22.70 | 31.46 | 207.41 |
| | middle | 7.52 | 29.29 | 15.57 | 52.38 |
| | low | 27.61 | 21.81 | 16.20 | 65.62 |
| | $\sum$ | 188.38 | 73.80 | 63.23 | 325.41 |

Table 5.4: Detailed view of contingency tables (a) and calculation of $\chi^2$ statistics (b) for domestic fires and population density. Zero valued cells are excluded from the analysis.

## 5.4 Conclusions

### 5.4.1 Capturing spatial and temporal aspects

This study demonstrates the use of contingency tables and $\chi^2$-based measures of associations as well established techniques used in traditional statistics to unveil bivariate relationships between spatio-temporal data. Careful conceptualisation of the spatial and temporal aspects during the pre-processing stage is necessary to enable the

(a) contingency table

| | | workplace density | | | |
|---|---|---|---|---|---|
| | | high | middle | low | $\sum$ |
| all fires density | high | 16 | 30 | 43 | 89 |
| | middle | 11 | 63 | 504 | 578 |
| | low | 2 | 49 | 1355 | 1406 |
| | $\sum$ | 29 | 142 | 1902 | 2073 |
| n-fires density | high | 4 | 18 | 57 | 79 |
| | middle | 14 | 45 | 270 | 329 |
| | low | 11 | 79 | 1575 | 1665 |
| | $\sum$ | 29 | 142 | 1902 | 2073 |
| d-fires density | high | 15 | 35 | 74 | 124 |
| | middle | 11 | 47 | 407 | 465 |
| | low | 3 | 60 | 1421 | 1484 |
| | $\sum$ | 29 | 142 | 1902 | 2073 |
| e-fires density | high | 15 | 26 | 83 | 124 |
| | middle | 8 | 51 | 485 | 544 |
| | low | 6 | 65 | 1334 | 1405 |
| | $\sum$ | 29 | 142 | 1902 | 2073 |

(b) $\chi^2$ test

| | | workplace density | | | |
|---|---|---|---|---|---|
| | | high | middle | low | $\sum$ |
| all fires density | high | 174.86 | 93.72 | 18.30 | 286.88 |
| | middle | 1.05 | 13.84 | 1.31 | 16.19 |
| | low | 15.87 | 23.24 | 3.27 | 42.39 |
| | $\sum$ | 191.78 | 130.80 | 22.88 | 345.46 |
| n-fires density | high | 7.58 | 29.28 | 3.31 | 40.17 |
| | middle | 19.19 | 22.39 | 3.36 | 44.94 |
| | low | 6.49 | 10.77 | 1.47 | 18.73 |
| | $\sum$ | 33.26 | 62.45 | 8.14 | 103.84 |
| d-fires density | high | 101.44 | 82.71 | 13.90 | 198.06 |
| | middle | 13.11 | 7.20 | 0.90 | 11.21 |
| | low | 15.19 | 17.07 | 2.59 | 34.85 |
| | $\sum$ | 119.74 | 106.99 | 17.40 | 244.13 |
| e-fires density | high | 101.44 | 36.08 | 8.32 | 145.84 |
| | middle | 0.02 | 5.06 | 0.40 | 5.48 |
| | low | 9.49 | 10.14 | 1.56 | 21.19 |
| | $\sum$ | 110.95 | 51.29 | 10.29 | 172.52 |

Table 5.5: Detailed view of contingency tables (a) and calculation of $\chi^2$ statistics (b) for domestic fires and workplace density. Zero valued cells are excluded from the analysis.

(a) contingency table

| | | density of households with adults | | | |
|---|---|---|---|---|---|
| | | high | middle | low | $\sum$ |
| all fires density | high | 53 | 32 | 2 | 87 |
| | middle | 116 | 377 | 53 | 546 |
| | low | 225 | 737 | 459 | 1421 |
| | $\sum$ | 394 | 1146 | 514 | 2054 |
| n-fires density | high | 40 | 35 | 2 | 77 |
| | middle | 84 | 191 | 43 | 318 |
| | low | 270 | 920 | 469 | 1659 |
| | $\sum$ | 394 | 1146 | 514 | 2054 |
| d-fires density | high | 55 | 56 | 6 | 117 |
| | middle | 104 | 285 | 64 | 453 |
| | low | 235 | 805 | 444 | 1484 |
| | $\sum$ | 394 | 1146 | 514 | 2054 |
| e-fires density | high | 51 | 67 | 4 | 122 |
| | middle | 100 | 339 | 83 | 522 |
| | low | 243 | 740 | 427 | 1410 |
| | $\sum$ | 394 | 1146 | 514 | 2054 |

(b) $\chi^2$ test

| | | density of households with adults | | | |
|---|---|---|---|---|---|
| | | high | middle | low | $\sum$ |
| all fires density | high | 79.01 | 5.64 | 17.95 | 102.60 |
| | middle | 1.21 | 17.19 | 51.19 | 69.59 |
| | low | 8.30 | 3.93 | 30.07 | 42.30 |
| | $\sum$ | 88.53 | 26.76 | 99.22 | 214.50 |
| n-fires density | high | 43.10 | 1.48 | 15.48 | 60.05 |
| | middle | 8.67 | 1.04 | 16.81 | 26.52 |
| | low | 7.31 | 0.03 | 6.98 | 14.33 |
| | $\sum$ | 59.08 | 2.55 | 39.27 | 100.90 |
| d-fires density | high | 47.23 | 1.32 | 18.51 | 67.06 |
| | middle | 3.37 | 4.12 | 21.49 | 28.98 |
| | low | 8.66 | 0.64 | 14.21 | 23.51 |
| | $\sum$ | 59.26 | 6.07 | 54.21 | 119.54 |
| e-fires density | high | 32.55 | 0.02 | 23.05 | 55.62 |
| | middle | 0.00 | 7.83 | 17.36 | 25.20 |
| | low | 2.79 | 2.77 | 15.59 | 21.15 |
| | $\sum$ | 35.34 | 10.62 | 56.00 | 101.96 |

Table 5.6: Detailed view of contingency tables (a) and calculation of $\chi^2$ statistics (b) for domestic fires and density of households with adults. Zero valued cells are excluded from the analysis.

method to be applied, especially in cases when the variables being studied are not categorical in nature. It includes a convenient representation of the spatial relationships and also a selection of suitable spatial and temporal analysis scales.

To embrace the geographical space, the proposed approach regards the variables being studied as grid layers, while proximity relationships are determined by a map overlay. The resolution of the analysis corresponds to the background data, which are considered sufficient for the given purpose. The variables being studied are represented by kernel density surfaces in order to minimise the uncertainties coming from rasterisation. Subsequent classification of the densities on the basis of natural breaks yields sensible categories reflecting the characteristics of the data distribution.

Careful pre-processing allows the method also to be used for temporal analysis. In this study the time dimension is incorporated by splitting the dataset according to selected temporal categories and performing the analysis on the data subsets separately. Prior knowledge about suitable time categories is necessary. Such an approach, although somewhat tedious, allows the way the relations between the analysed data change in time to be unveiled.

### 5.4.2 Type of knowledge discovered

The analysis of contingency tables is useful for finding correlations in the data that are represented as categorical. The unveiled relations are bivariate, as contingency tables usually sort the data according to two variables at time, separated from remaining attributes.

The variables are tested against the hypothesis of independence, which is or is not rejected on the basis of established statistical measures. The relationships can also be quantified to express the strength of the correlation. In this way the relations between several variables can be statistically compared in order to identify the most significant ones.

Contingency tables also provide a deep insight into the relations between the particular data categories, for example, cells invoking high values of the $\chi^2$ measure, and thus the reasons for rejecting the hypothesis of independence between the variables can be identified. A comparison of the observed and expected cell counts unveils further details in terms of positive or negative correlation.

### 5.4.3 Weaknesses

Contingency tables are used for a bivariate analysis. Such an analysis may be of interest to the user, particularly if it enables the strength of the correlations to be quantified. However, the underlying phenomenon is usually more complex. The user should be aware that a purely bivariate analysis can be misleading and should be complemented by a multivariate approach.

Contingency tables can be applied to quantify relationships between categorical variables. If the data being analysed are not categorical in nature, suitable conceptualisation and categorisation, which may result in an information loss, is necessary. As demonstrated in this study, the representation of the studied variables as a density surface divided into three categories related to a grid cell obscures insight into the details contained in the original data. However, if attention is paid to the pre-processing step and the categories selected correspond to the nature of the data distribution, the analysis of contingency tables provides an effective characterisation of the data being studied in terms of the significance of the relationships.

A further issue with the application of contingency tables to geographical data concerns an interactive connection with a map, which is necessary for the exploration of spatial data. The analysis alone offers flexibility in data exploration by modification of the categories being studied; however, the possibility of highlighting the spatial locations of the categories of interest effectively during the analysis is missing.

The analysis of the contingency tables is based on statistical testing and numerical evaluation. Insufficient utilisation of the potential of visualisation makes the interpretation of the relationships and also the presentation of the results difficult.

### 5.4.4 Requirements for the user

Data analysis using contingency tables represents one of the fundamental methods based on solid statistical grounds. It requires knowledge of basic statistical concepts, but its further application is straightforward. The interpretation of the results is also relatively easy, which minimises the risk of misunderstandings in the communication between the parties involved in the process of knowledge discovery.

Nowadays, the statistical analysis of data using contingency tables is supported by numerous data management software packages. MS Excel is used in this study as it is one of the most common spreadsheet processors, and may therefore be familiar

to potential users beforehand. It also provides flexible tools for the further analysis and display of the data.

The application of contingency tables to geographical data requires extensive preprocessing before the actual analysis can be performed. Thus, the role of a GIS expert capable of data analysis on both the conceptual and technical levels is crucial for a potential future user in order for the method presented to be employed successfully.

# Chapter 6

# Point pattern analysis

The following article forms the basis of this chapter. The author performed the analysis, interpreted and summarised the results, and played a leading role in writing the paper.

Špatenková, O., Stein, A. (2009) Identifying factors of influence in the spatial distribution of domestic fires. To appear in International Journal of Geographical Information Science.

## 6.1 Method

Domestic fires as observed point locations distributed over a region of space form a spatial point pattern, for which statistical analysis methods are described in the literature, e.g. [Bailey and Gatrell, 1995; Diggle, 2003; O'Sullivan and Unwin, 2003]. Point pattern analysis aims to characterise the properties of the point pattern by detecting any systematics, either regularity or aggregation (clustering), in the distribution of points.

The analysis usually begins with tests for random distribution, for which the density of the point pattern does not vary over the bounded region, and there are no interactions among the points. Such a pattern is formally defined as complete spatial randomness (CSR) by the following criteria: (i) the number of events in planar region $A$ with area $|A|$ follows a homogeneous Poisson distribution with mean $\lambda|A|$, where $\lambda$ is the constant density; (ii) given $n$ events $x_i$ in a region $A$, the $x_i$ are an independent

random sample from the uniform distribution on $A$ [Diggle, 2003].

Various summary statistics can be derived illustrating first-order effects, i.e. differences in the number of points per a unit area over the study region, or second-order effects, i.e. dependence relationships between the points. Density is a measure of first-order effects. Its estimation is usually based on kernel methods [Silverman, 1986; Bowman and Azzalini, 1997], when the density surface is composed of kernels – bivariate probability density functions symmetric about the origin placed at each point location. Neighbouring points contribute to the density estimation; their influence decreases with the distance from their kernel centre. The usual estimator of the kernel density $\lambda$ at location $u$ is

$$\lambda(u) = \sum_{i=1}^{n} \kappa_h(u - x_i),$$

where $\kappa$ is the Gaussian kernel function with a bandwidth $h$ and $n$ is the number of observed points $x_i$ in the study region. The kernel search radius, also called a bandwidth, determines the smoothness of the density surface – too large a bandwidth results in a general picture approaching the average of the study area, while a small bandwidth focuses on the individual data records and reflects any slight variations. Experimentation with the bandwith setting is necessary to derive an optimal density surface for a given application [Krisp and Špatenková, 2009]. Well-chosen density plots provide a better insight into the data than a dot map, where exploration becomes difficult as a result of overprinting.

Nearest neighbour distances describe the dependence relationships for small-scale interactions. They are defined as the distance from the $i$th point to the nearest other point in the bounded region of interest. The empirical cumulative probability distribution function $\hat{G}$ for the nearest neighbour distances provides an effective summary of the point pattern:

$$\hat{G}(w) = \frac{\sum_{w_i \leq w} 1}{n},$$

where $w_i$ is a direct nearest neighbour distance for the $i$th point and $n$ is the number of points in the study region.

As the observed point pattern is usually part of a larger region, where the distribution of points outside the study area is unknown, the interaction between these points cannot be properly accounted for, causing edge effects [Diggle, 2003]. One of the methods for adjustment, which is simple but effective, consists of reducing the

sample by the buffer defined around the boundary. Points which fall inside the buffer are not used for the analysis directly, but unveil the distribution behind the reduced study region.

The $\hat{G}$ plot reflects the spatial distribution of the pattern. While an excess of nearest neighbours at short distances indicates clustering in the data, an excess of long-distance neighbours refers to regularity. For easier interpretation it is suitable to compare the $\hat{G}$-function with the theoretical curve for CSR, which is (ignoring the edge effects):

$$G(w) = 1 - exp(-\lambda \pi w^2).$$

Monte Carlo simulations of the CSR process provide an insight into the statistical significance of the difference between the empirical and theoretical curves. The CSR with the same density as the data pattern is simulated in 99 realisations, whereas empirical cumulative probability distribution functions for the nearest neighbour distances are constructed for each of them. Maximum and minimum values form simulation envelopes, which can be interpreted as critical values for testing the null hypothesis of randomness.

The analysis of the nearest neighbour distances between domestic fires and additional point patterns, representing background building and population characteristics, provides an insight into the relationships in the data. If the short distance nearest neighbours exceed the expectation for random distribution, a correlation between the datasets can be assumed, unveiling the process underlying the distribution of domestic fires. A comparison of the $\hat{G}$-function plots also indicates the importance of particular exploratory variables.

A spatial point process is a stochastic mechanism that generates a point pattern in the study region. Fitting a process of domestic fires (DF process) with the density function that reflects the spatial distribution of the studied influences enables the dependence of domestic fires on exploratory variables to be modelled. As the data exhibit a stochastic dependence between the points, a non-stationary Strauss process [Strauss, 1975; Kelly and Ripley, 1976] is selected as a flexible model for modelling pairwise interactions. The conditional density of the Strauss process is

$$\lambda(u, x) = \beta(u) \cdot \gamma^{t(u,x)},$$

where $\beta(u)$ is the density at location $u$, $t(u, x)$ is the number of points $x$ that lie within a distance $r$ of $u$, and $\gamma$ is an interaction parameter that controls the strength of the

interaction between points. For the special case when $\gamma = 1$ the model reduces to the homogeneous Poisson process with constant density $\beta$; the case that $\gamma = 0$ gives a simple inhibition process, whereas for $\gamma > 1$ the model corresponds to a clustered process. The dependence on exploratory variables is expressed as a density, which is a loglinear function of covariates:

$$log\beta(u) = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_n c_n,$$

where the $c_i$ are the explanatory variables and $\beta_i$ are parameters to be fitted.

Modelling the spatial point pattern aims to find a suitable representation of the data corresponding to observed relationships in an iterative process. The models are inspected according to several criteria. The overall goodness of fit for the Strauss models is assessed using simulation envelopes of summary functions [Diggle, 2003; Mattfeldt et al. 2007; Baddeley, 2008]. The $K$-function is more effective than measures based on nearest neighbour distances, as it provides a summary of the spatial pattern over a wide range of scales. It is defined as the expected number of other points of the process lying within a distance $d$ of a typical point of the process, divided by the density $\lambda$. A suitable estimate of the $K$-function given by [Ripley, 1976] is:

$$\hat{K}(d) = \frac{R}{n^2} \sum_{i=1}^{n} \sum_{j \neq i} I_d(d_{ij}),$$

where $n$ is the number of points in the study region with area $R$, $d_{ij}$ is the distance between the $i$th and $j$th points, and $I_d(d_{ij})$ is an indicator function, which is 1 if $d_{ij} \leq d$ and 0 otherwise. The $\hat{K}(d)$ adjusted for inhomogeneity becomes:

$$\hat{K}_I(d, \lambda) = R^{-1} \sum_{i=1}^{n} \sum_{j \neq i} \frac{I_d(d_{ij})}{\lambda(x_i)\lambda(x_j)}.$$

The reduced sample method is used to adjust the estimate for edge corrections. The test of the goodness of fit considers global envelopes of the $\hat{K}_I$-function calculated for each of the realisations of simulated models. These represent the largest absolute differences between the simulated and estimated theoretical curves over the entire distance interval. The test requires 19 simulations in order to achieve a 5% significance level [Baddeley and Turner, 2005].

The Akaike Information Criterion AIC [Akaike, 1974] is used for model selection. It represents a goodness of fit when the number of estimated parameters and the

number of observations are being considered at the same time. A model with the lowest AIC value thus reflects the best trade-off between bias and variance.

The analysis was conducted using spatstat, an R package designed for analysing spatial point patterns [Baddeley and Turner, 2005; Baddeley, 2008].

## 6.2 Data pre-processing

The spatial point pattern analysis methods take advantage of the original representation of the incident dataset. Background attributes form additional point patterns associated with the incidents during the analysis, which simplifies the pre-processing. Census data, which are supplied in grid format, need to be converted into a point pattern by attaching the attributes to the central points of the particular grid cells. The resolution of the grid, 250 $\times$ 250 metres, is considered sufficient to justify this step in the present application.

Besides the spatial location, all fires are also characterised by their time of occurrence. A daily scale which exhibits the most significant variations in the density is selected for the analysis. The three categories, representing n-fires (occurring between 1 a.m. and 8 a.m.), d-fires (occurring between 8 a.m. and 5 p.m.), and e-fires (occurring between 5 p.m. and 1 a.m.) are treated individually and the results are compared to unveil temporal changes in the distribution of domestic fires.

## 6.3 Results

First the kernel density of the domestic fires is analysed. A suitable smoothing is achieved experimentally with a kernel bandwidth of 200 m. Figure 6.1 illustrates temporal variations in the distribution of domestic fires over the three daytime categories. The average density is highest in the evenings and lowest at night. The fire density is also analysed in relation to the remaining temporal attributes. Table 6.1 shows the average density of domestic fires described in Figure 3.2 (b) according to the daytime categories. An increase in the fire density is observed on Friday and Saturday evenings. The patterns differ significantly from random distribution, unveiling two hotspots in the centre of Helsinki.

Distance statistics characterise the distribution of domestic fires in terms of spatial dependence. Figure 6.2 shows the $\hat{K}$-functions plotted against the theoretical value

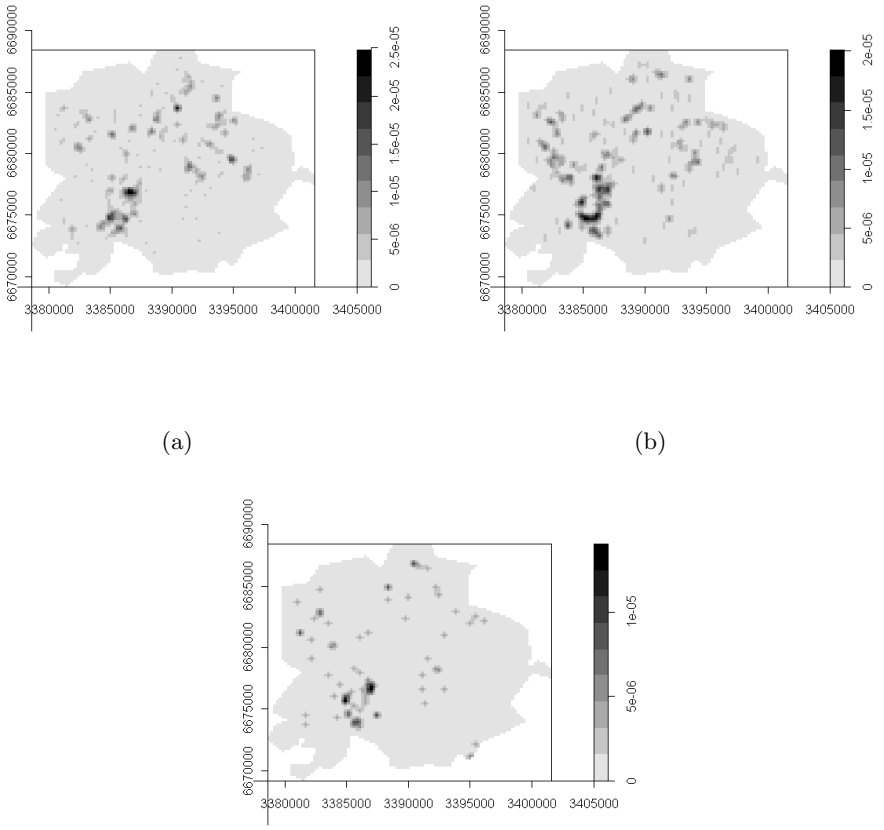
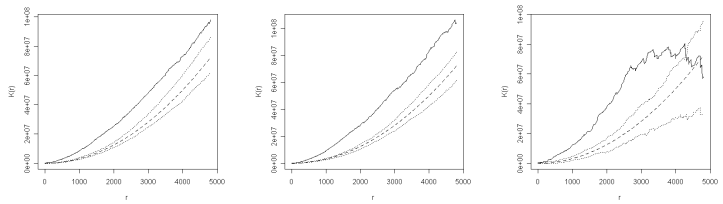



Figure 6.1: Density plots of domestic fires (kernel bandwidth 200 m): (a) e-fires; (b) d-fires; (c) n-fires.

| Weekday | e-fires | d-time fires | n-fires |
|---|---|---|---|
| Monday | 1.09 | 1.35 | 0.224 |
| Tuesday | 1.12 | 1.28 | 0.513 |
| Wednesday | 0.96 | 1.25 | 0.192 |
| Thursday | 0.74 | 1.25 | 0.449 |
| Friday | **1.57** | 0.74 | 0.224 |
| Saturday | **1.35** | 1.22 | 0.256 |
| Sunday | 1.09 | 1.06 | 0.545 |

Table 6.1: Average density ($\times 10^{-7}$) of daytime categories of domestic fires per $m^2$ studied for different days of the week. Notice the increased values during Friday and Saturday evenings.



(a) (b) (c)

Figure 6.2: $\hat{K}$-functions for domestic fires (solid line) with theoretical values for random distribution (dashed line) and simulation envelopes (dotted line): (a) e-fires; (b) d-fires; (c) n-fires.

for CSR and its simulation envelopes. The empirical curve lying above the theoretical line and upper simulation envelope indicate clustering in the data.

In the next step, nearest neighbour distance analysis is applied to unveil the dependence of the domestic fires on the background environment. The relationships observed do not differ significantly between the particular daytime categories, indicating that temporal aspects have mostly first-order effects (density distribution) rather than second-order effects.

Figure 6.3 shows the $\hat{G}$-function plots of the distribution of domestic fires in relation to socio-economic attributes. Naturally, population and workplace densities are identified as important influences. The empirical $\hat{G}$-functions lie above the theoretical

lines at short distances in the cases of the high- and middle-density categories, indicating that areas with high and middle densities of inhabitants and workplaces are affected by domestic fires more than would be expected from randomness. Next, areas with a high ratio of households with adults correlate with a high level of occurrence of fires. Households with children, in contrast, suffer less from domestic fires. As the empirical $\hat{G}$-function falls between the simulation envelopes in the case of pensioners' households, the null hypothesis of random distribution is not rejected. Low incomes of the inhabitants are also identified as relevant aspects. Further, the middle and low education and middle and high unemployment categories show an influence on the distribution of domestic fires.

The distribution of domestic fires also exhibits a dependence on the building attributes under study (Figure 6.4). There exists a correlation between housing and work building types, while fires in leisure buildings are random. Compared to new buildings, middle-aged and old buildings slightly increase the fire risk.

As a third step, the study aims to model the dependence of domestic fires on the factors of influence under study by fitting a non-stationary Strauss process with its intensity being a loglinear function of the explanatory variables. A model involving all the variables studied is a starting point. In order to simplify the model, particular variables are excluded in an iterative process with the aim of improving the goodness of fit. The reduced models are compared on the basis of the AIC measure. As the AIC values depend on the amount of fires in particular categories (247, 254, and 75 for $e$-, $d$-, and $n$-fires, respectively), it cannot be used as a mutual quality measure of the models, but it does serve for model selection. Table 6.2 lists the estimated parameters and the AIC values for the full and selected reduced models.

The full model considers the whole dataset of domestic fires and involves all the influences that were studied as covariates. The value of the interaction parameter $\gamma > 1$ indicates a clustered $DF$ process. It means that locations close to each other are more likely to have the same fire outbreak pattern than locations further away. Other parameters suggest the effect of particular variables on the modelled density in terms of a positive or negative contribution and also indicate the changes in the variable impact within a category, or over time. The population density, building type, and income categories emerge as important influences.

The temporal aspect is analysed by fitting the full models to $e$-, $d$-, and $n$-fires separately. The differences observed between the models suggest keeping the distinc-

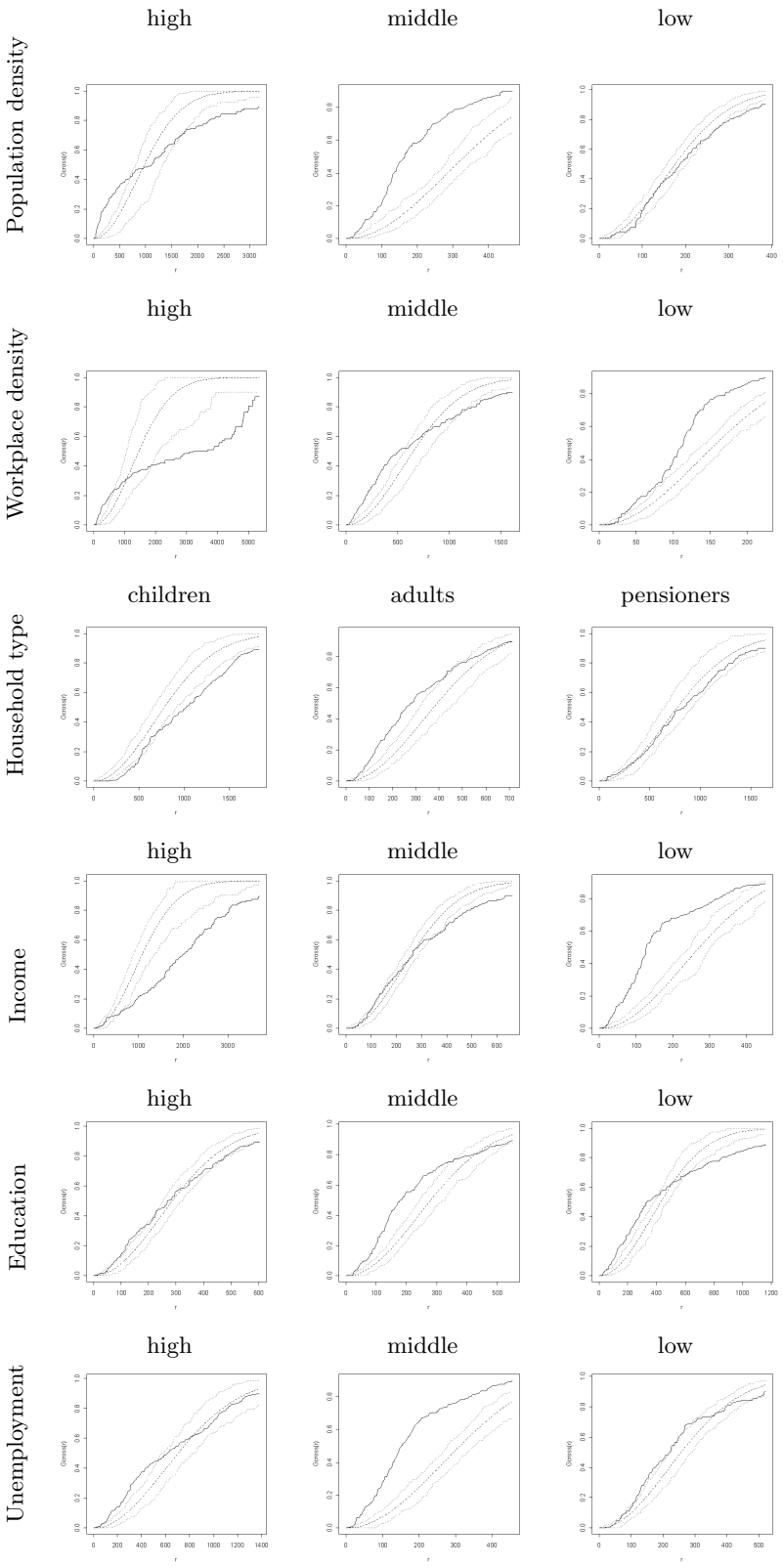



Figure 6.3: $\hat{G}$-functions showing relations between domestic fires and population attributes (solid line) with theoretical values for random distribution (dashed line) and simulation envelopes (dotted line).

| | Full model | | | | Reduced model | | |
|---|---|---|---|---|---|---|---|
| | all | e-fires | d-fires | n-fires | e-fires | d-fires | n-fires |
| Intercept | -14.64 | -15.62 | -15.40 | -16.44 | -15.66 | -15.48 | -16.52 |
| Building type | | | | | | | |
| leisure | 3076* | 13733* | -11912* | 6976 | -1288* | -4726* | -3201* |
| work | 14747* | 28324* | -1943* | 13186 | 8325* | 9815* | 11041* |
| housing | 6564* | 18845* | -85996* | 4229 | 387* | 9911* | 916* |
| Year of construction | | | | | | | |
| $<$ 1945 | -20187 | -17715 | 9411 | -6706 | | | |
| 1945 – 85 | -18134 | -16160 | 9502 | -1295 | | | |
| $\geq$ 1985 | -17232 | -14585 | 10331 | -7345 | | | |
| Pop. dens. | 44096* | 65309* | 3120 | 295428* | 47615* | 41245* | 164939* |
| Workplace dens. | 4836 | 34152 | 20047 | 94447 | | | |
| Household type | | | | | | | |
| with children | -115703* | -150145 | -138820 | -19453 | | | |
| with adults | -116554* | -29206 | 4585 | -49559 | | | |
| with pensioners | 26283* | 60171 | -30355 | -55760 | | | |
| Education | | | | | | | |
| high | -58095 | -62901 | -9189 | -642960 | | | |
| middle | -98179 | -108710 | -45806 | -604783 | | | |
| low | -93342 | -96557 | -35791 | -577164 | | | |
| Unemployment | | | | | | | |
| high | 108699 | 114711 | 129233 | -178899 | | | |
| middle | 149277 | 198616 | 146996 | -288706 | | | |
| low | 125730 | 173532 | 120486 | -357057 | | | |
| Income | | | | | | | |
| high | 83965* | 145461 | -16625* | 902565* | 176793* | 47873* | 26992* |
| middle | 24609* | 27665 | -46491* | 830104* | 78813* | 66578* | 32466* |
| low | 98306* | 80551 | 52841* | 962378* | 164037* | 154310* | 115466* |
| $\gamma$ | 1.32 | 1.59 | 1.48 | 1.88 | 1.59 | 1.50 | 1.85 |
| AIC | 15331.3 | 7007.3 | 7208.7 | 2321.1 | 6997.8 | 7192.2 | 2309.2 |
| Nr. of observations | 576 | 247 | 254 | 75 | 247 | 254 | 75 |

Table 6.2: Parameter values for the selected model (significant parameters marked with *).

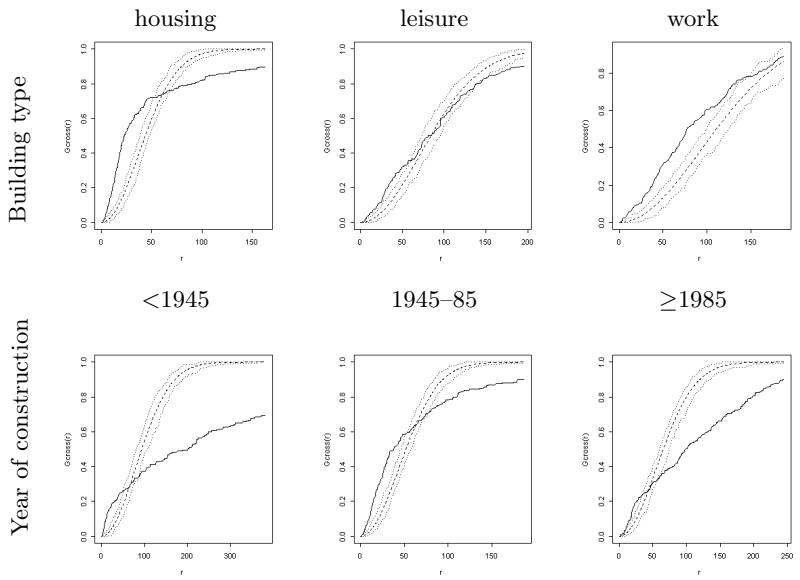
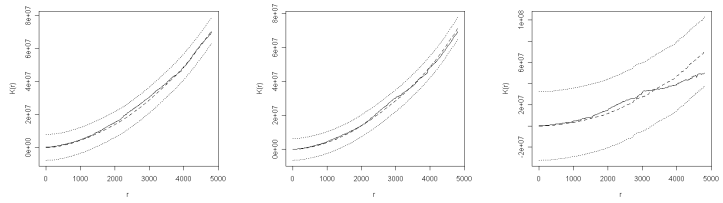



Figure 6.4: $\hat{G}$-functions showing relations between domestic fires and building attributes (solid line) with theoretical values for random distribution (dashed line) and simulation envelopes (dotted line).

tion and further treating the temporal models individually. The complexity of the models is reduced by excluding the variables step by step, while comparing the AIC values. The selected reduced models show a decrease in the AIC values with 9.5 for the *e*-fires, 16.5 for the *d*-fires, and 11.9 for the *n*-fires. The overall goodness of fit is tested on the basis of simulations. The $\hat{K}_I$-function estimated from the original pattern falls between the simulation envelopes representing 5% critical bands for the inhomogeneous Strauss process fitted to the data (Figure 6.5).

The temporal models that are derived contain the same variables, which are building types, population density, and the average incomes of the inhabitants. However, the parameter values and thus also the density estimates of these models vary (Figure 6.6). All the models unveil the hotspots in the centre of Helsinki. The interaction parameter $\gamma$ indicates clustering in the distribution of fires, which is strongest for *n*-fires. In the case of *e*-fires, additional significant clusters are found to the northwest of the city centre. The lowest value of $\gamma$ corresponds to fewer interactions in the distribution of *d*-fires around the study area. Parameter values indicate temporal variations in their influence, as the effect of population density is strongest at nights,



(a) (b) (c)

Figure 6.5: Goodness of fit of the selected best temporal models based on the $\hat{K}_I$-functions for the observed domestic fires (solid line), estimated theoretical value (dashed line), and global simulation envelopes (dotted line): (a) e-fires; (b) d-fires; (c) n-fires.

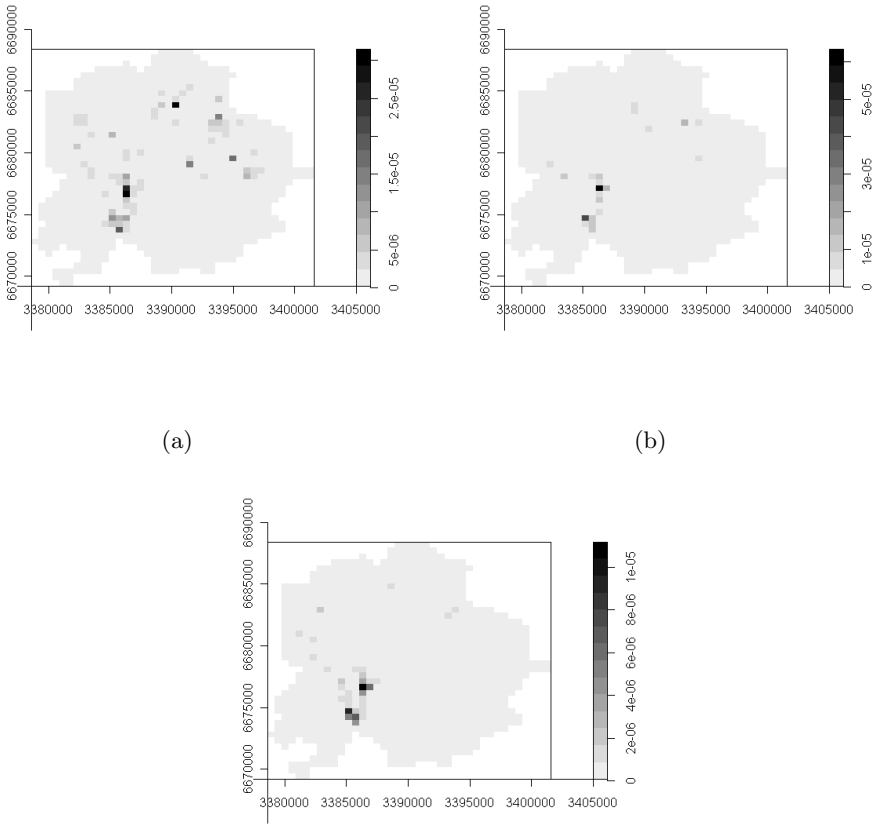
when the population model based on the permanent addresses corresponds best to the true distribution of the inhabitants. The low income category is connected mainly to *e*-fires, whereas inhabitants with medium incomes affect *d*- and *e*-fires. Building types are also an important influencing factor. In contrast to work buildings, which exhibit an even influence throughout the day, housing buildings are strongly connected to *d*- and *n*-fires. Buildings classified as leisure buildings have a negative effect on the distribution of domestic fires.

## 6.4 Conclusions

### 6.4.1 Capturing spatial and temporal aspects

Point pattern analysis methods have been developed specifically for dealing with spatial data represented as points. The use of point pattern analysis allows the level of detail offered by the original data to be preserved and the necessary pre-processing to be minimised. In contrast to, for example, lattice methods, which handle aggregated data, point pattern analysis avoids an ambiguous definition of a lattice scale and therefore provides more accurate results.

The temporal aspect is incorporated into the analysis by splitting the dataset into pre-defined time categories. These are treated separately and the results are compared to discover variations on the temporal scale. Such an approach is, however, slightly cumbersome, as it hampers the modification of the selected categories or zooming in




(a) (b)

(c)

Figure 6.6: Fitted density function of the selected best temporal models including population income, building types and population density as covariates: (a) e-fires; (b) d-fires; (c) n-fires.

on other scales during the analysis.

### 6.4.2 Type of knowledge discovered

Point pattern analysis enables any systematic features in the distribution of domestic fires to be detected in terms of difference from randomness. First-order effects, expressed as a kernel density, indicate variations in the distribution over the study region. Second-order effects as applied in this study discover the spatial dependence between the incidents and individual factors of influence. A comparison of the $\hat{G}$-function plots indicates significant influences, while providing an insight into the aggregation scales.

In contrast to the distance functions, modelling the $DF$ process provides a multivariate insight into the relations, as it considers all the variables simultaneously. It also enables the influence of the variables through the estimated parameters, including differences on the temporal scale, to be quantified. Goodness of fit and model selection measures help in identifying the most important factors of influence.

### 6.4.3 Weaknesses

The methods used in a point pattern analysis require the phenomenon being studied to be represented as a point pattern. This is a natural choice for the incident records. Background census attributes are, however, originally represented as grid cells. Careful conceptualisation and pre-processing are necessary. In the study presented, these steps include the representation of the grid cells as a point pattern, with the spatial resolution being taken into account. In addition, the background datasets are split into various categories that may have vague boundaries between them. As such a categorisation may affect the uncertainty of the results, special attention must be paid to the conceptualisation and pre-processing steps.

Kernel density surfaces are used for the effective representation of the point data distribution. The result of the kernel density estimations depends on a number of parameters, while the choice of the kernel bandwidth is essential. The selection of the bandwidth may be subject to formal analysis, but it is often more an art than a science [Krisp and Špatenková, 2009]. The application, spatial resolution, and purpose of the analysis must be carefully considered in order for a suitable value of the bandwidth to be found.

The analysis of the nearest neighbour distribution on the basis of the $\hat{K}$- and

$\hat{G}$-function plots is valuable in order to detect clustering or regularity in the data and correlations between two point patterns. These represent univariate and bivariate relationships. A careful analysis should continue by also exploring multivariate relationships in order to describe the complex underlying process and facilitate their understanding.

The $\hat{K}$- and $\hat{G}$-function plots are qualitative measures that indicate the relationships in the data. Modelling the underlying process on the basis of spatial covariates enables the relationships to be quantified. Note that the point process represents a global model in this case.

### 6.4.4 Requirements for the user

The point pattern analysis comprises advanced methods, whose application requires an insight into spatial statistics and regression analysis from the user. Visual representation of the results, such as density maps or distance function plots, facilitates the interpretation and also provides a suitable tool for communication with domain experts and other involved parties.

The spatstat [Baddeley and Turner, 2005; Baddeley, 2008] package for the statistical analysis of spatial data in R or S-PLUS is used in this study. The current version is capable of dealing with patterns of points in the plane and supports the manipulation of point patterns and exploratory data analysis, as well as model-fitting and the simulation of point processes. It is therefore found suitable for the purposes of this study.

# Chapter 7

# Geographically weighted regression

## 7.1 Method

The spatial processes we aim to investigate are often non-stationary, which means that spatial variations exist in relationships over space. When global statistics, which emphasise similarities across space, are applied to such processes, they produce nothing more than misleading summaries of the complex multivariate relationships. In order to unveil a great deal of information about the process, there is a need to examine detailed spatial variations in relationships. A number of local methods have been proposed as a response to this need. Geographically weighted regression (GWR), developed by Fotheringham et al., represents a method that allows the local analysis of relationships in spatial datasets [Fotheringham et al., 2002]. This technique is based on the traditional regression framework and incorporates local spatial relationships in an intuitive and explicit manner.

Regression analysis is a method that aims to model the relationship between one or more dependent (also called response) variables and several independent (or predictor) variables. In a linear form, the regression model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \epsilon_i, \tag{7.1}$$

where the influence of $n$ independent variables $x$ on the dependent variable $y$ at point $i$ is represented by particular parameters $\beta$, and $\epsilon_i$ is a random error with normal

distribution. In the case of a global model, the parameters $\beta$ are constant over space, resulting in the same response to the same stimulus in all parts of the study region.

GWR extends this framework by allowing variations of the parameters $\beta$ as a function of location represented by the coordinates $(u_i, v_i)$:

$$y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_{i1} + \beta_2(u_i, v_i)x_{i2} + \cdots + \beta_n(u_i, v_i)x_{in} + \epsilon_i. \tag{7.2}$$

The calibration of the GWR model is problematic, as there are more unknown parameters than observed variables. However, estimates with only a small amount of bias can be provided, as the parameters are not assumed to be random. According to Tobler's first law of geography, all objects are related to each other, but closer objects are more closely related than distant ones. We can therefore assume some degree of spatial auto-correlation in the parameter values being estimated, and approximate the GWR model locally by performing a global regression using a subset of nearby points. Assuming that observed data near to the location being estimated should have more of an influence, the estimate of $\beta$ is a weighted least-squares estimator, for which the weight matrix varies according to the location of point $i$:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = (\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i)\boldsymbol{Y}, \tag{7.3}$$

where the bold type denotes a matrix and $\boldsymbol{W}(u_i, v_i)$ is an $n \times n$ matrix, whose off-diagonal elements are zero and whose diagonal elements denote the geographical weighting of each of the $n$ observed data for regression point $i$.

A typical weighting function fits a spatial kernel to the data as shown in Figure 7.1. The kernel is centred at a given regression point, whereas the weight decreases with an increase in the distance from this point. Gaussian shape is commonly used for the weighting function. The spatial kernels can be fixed in terms of their shape and magnitude over space, or they can be allowed to vary spatially, with the bandwidth being smaller in regions with a high density of data points and larger where the density of data points is low, as shown in Figure 7.2. In practice, the results of GWR are relatively insensitive to the choice of weighting function; however, they are sensitive to the degree of its distance-decay [Fotheringham, 2009]. The selection of an optimal bandwidth is always a trade-off between bias and variance, as with too small a bandwidth the parameters will depend on observations in close proximity to

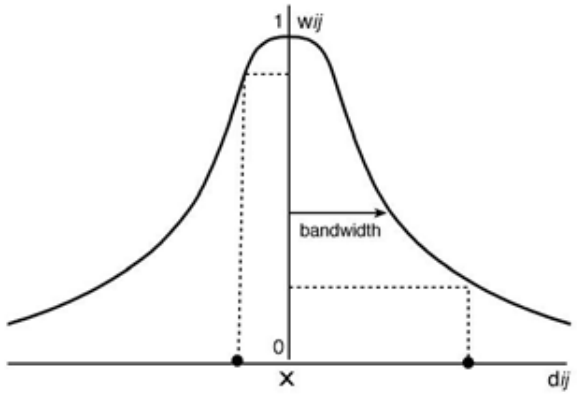
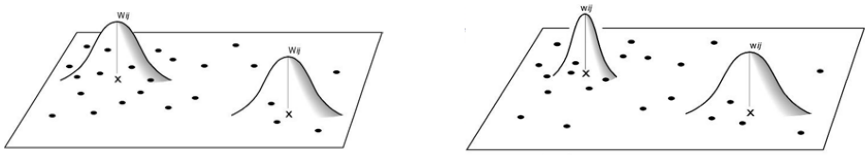



Figure 7.1: A spatial weighting function [Fotheringham et al., 2002]. A cross denotes a regression point, dots represent data points, $w_{ij}$ is the weight of data point $j$ at regression point $i$, and $d_{ij}$ is the distance between regression point $i$ and data point $j$.

the regression point increasing the variance, whereas if the bandwidth is too large, GWR starts to resemble a global model. The Akaike Information Criterion (AIC) discussed in the previous chapter can be used for selection of the optimal bandwidth. In contrast to other approaches, AIC has the advantage of being more general and it can also be used to compare the goodness of fit between a global model and GWR, while considering different degrees of freedom in the models.

In order to understand various aspects of model performance, it is useful to compute standard regression diagnostics. A measure representing how well the model fits the data in known as $r^2$. It represents the percentage of variance in the response variable accounted for by a variance in the model, while the remaining variance is accounted for by other things not included in the model.

Local standard errors represent a measure of uncertainty of the parameter estimates. If the parameter estimate is divided by its standard error, we obtain a $t$ value. The $t$ value enables the probability of a parameter being zero to be determined, for which changes in a predictor variable have no effect on the estimated variable. $T$ values therefore imply the significance of the relationship between the predictor and response variables. In an exploratory manner, $t$ values are used to highlight parts



(a) (b)

Figure 7.2: Weighting schemes [Fotheringham et al., 2002]: (a) fixed, (b) spatially adaptive. A cross denotes a regression point, dots represent data points.

of the study area where interesting relationships appear to be occurring. Absolute $t$ values greater than 1.96 and 2.58 are often used to determine the significance levels of 95% and 99%, respectively [Fotheringham et al., 2002].

Residuals represent the difference between the observed and the predicted value of a response variable. Positive residual values indicate an underestimate by the model, whereas negative values indicate an overestimate. Residuals can be standardised to have a mean of 0 and a variance of 1, which helps in the assessment of their significance. Residuals can also be mapped to provide information about the spatial stationarity of the process. Some degree of positive spatial autocorrelation can be found in most spatial applications. As global models do not account for that, their positive residuals also have positive residuals as neighbours and vice versa. This leads to the underestimation of the standard errors of the parameter estimates and potential problems with inference. Spatial regression techniques [Fotheringham et al., 2000] are often applied as a solution to this problem. In order to account for spatial autocorrelation, they consider an additional explanatory variable reflecting the values of the dependent variable in the neighbourhood. GWR, in contrast, models the non-stationarity directly by allowing the parameters to vary spatially. The residuals from GWR are therefore not spatially autocorrelated. In addition, they are usually much lower than those of the global models.

The significance of the spatial variability in the local parameter estimates is formally examined by conducting a Monte Carlo test [Hope, 1968]. The results indicate the variables exhibiting significant spatial non-stationarity, which cannot be accounted

for in the global model.

A software package for GWR (release 3.4.3) developed by the authors of the geographical weighting idea was used to perform the analysis.

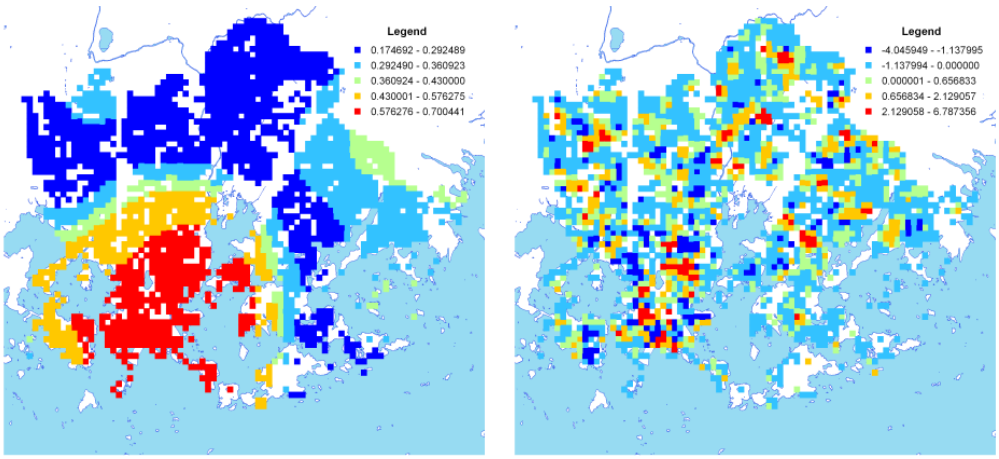
## 7.2 Data pre-processing

The distribution of domestic fires, as a dependent variable, is represented by a kernel density surface. A kernel bandwidth of 250 m is selected for this application. The density surface is formed of grid cells corresponding to the form of census records. The socio-economic population attributes and building characteristics constitute independent variables, and are related on the basis of a map overlay. In this case, the building attributes reflect the average age of the buildings and a prevailing building type in the grid cell. The fire dataset is further split according to the hourly scale for temporal analysis. The regression model is therefore constructed for all fires, n-fires (1 a.m. – 8 a.m.), d-fires (8 a.m. – 5 p.m.), and e-fires (5 p.m. – 1 a.m.).

## 7.3 Results

The analysis begins with the full dataset including all fires. A fitted global model estimates parameter values for 10 predictor variables, while 5 of them, i.e. the building type, population density, density of workplaces, ratio of households with children and unemployment rate, are according to $t$ values significantly different from zero; see Table 7.1. Cut-off values of 1.96 and 2.58 determine the significance levels of 95% and 99%, respectively. The global parameter values show a positive association of the significant variables, except households with children, which exhibit a negative association. The AIC for this model is 11855.1, the adjusted $r^2$ is 0.43, and the global coefficient of determination is 0.43, which means that 57% of the variation in the estimated density of fires is not accounted for in the model.

The AIC value of the corresponding GWR model is 11656.0, which means a decrease of 199.9 compared to the global model; see Table 7.2. A higher value of the coefficient of determination indicates that 51% of the variation in the density of fires is explained by the model. It means that the local model significantly improves the global estimate. As the variables are represented in the form of grid cells regularly distributed over the study area, a fixed kernel method is selected to determine the




(a) (b)

Figure 7.3: GWR results for a full model of all fires: (a) Local R-Squared; (b) Standardised residuals. The red-green-blue colour scale indicates where the GWR provides a better fit compared to the global models according to local $r^2$ (red-green areas) and distribution of positive-negative standardised residuals.

local neighbourhood. The lowest AIC value is reached at a bandwidth of 1909.7 m. The significance of the spatial variability of the parameters is based on Monte Carlo simulations. Six out of ten parameters exhibit significant spatial variability compared to the global model.

The main output from GWR, a set of local parameter estimates and associated diagnostics, is mapped in Figures 7.3–7.6. The colour scheme in Figure 7.3 (a) indicates where the GWR model provides a better fit than the global model (reddish areas). The standardised residuals in Figure 7.3 (b) show no distinct autocorrelation. The distribution of the significant parameter values is illustrated in Figures 7.4–7.6. The colour scale from blue through green to red indicates changes in the parameter values (negative to positive), while the grey colour masks zero-valued parameters based on local $t$ values. We can observe a significant spatial variation in the parameter values. Households with pensioners, for example, exhibit a positive association with the density of domestic fires in some areas and a negative in others. However, the values are significant only in small parts of the study area.

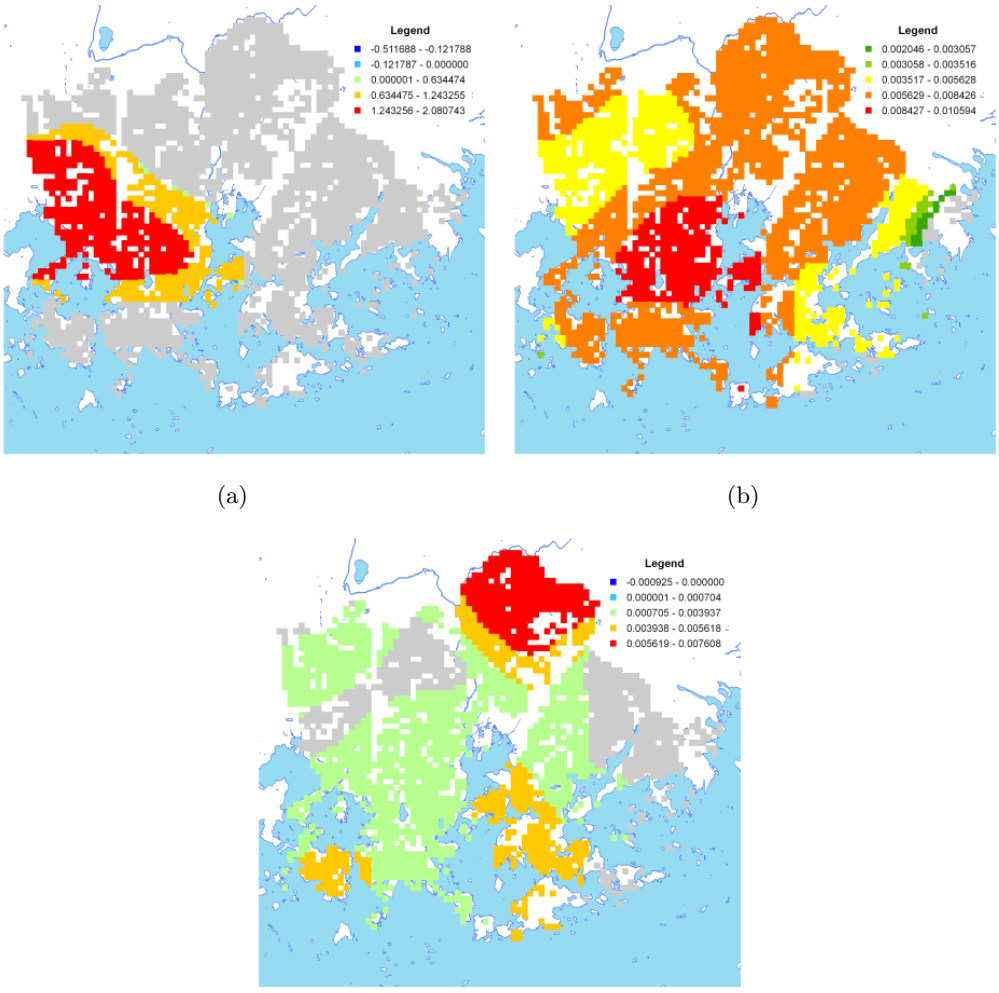
In the next step, the regression is applied to the three selected temporal data

| | All fires | | | D-fires | | | E-fires | | | N-fires | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AIC | 11855.9 | | | 9076.5 | | | 9549.0 | | | 7371.8 | | |
| Adjusted $r^2$ | 0.43 | | | 0.29 | | | 0.27 | | | 0.26 | | |
| Coeff. of determination | 0.43 | | | 0.29 | | | 0.28 | | | 0.26 | | |
| Parameter | Estimate | St Err | T | Estimate | St Err | T | Estimate | St Err | T | Estimate | St Err | T |
| Intercept | 0.5e-1 | 9.2e-1 | 0.1 | -1.8e-1 | 4.8e-1 | -0.4 | 2.3e-1 | 5.4e-1 | 0.4 | -0.4e-1 | 3.2e-1 | -0.1 |
| Building age | -0.6e-4 | 4.5e-4 | -0.1 | 0.1e-4 | 2.4e-4 | 0.1 | -0.5e-4 | 2.6e-4 | -0.2 | 0.2e-4 | 1.6e-4 | 0.1 |
| Building type | **5.7e-1** | **1.6e-1** | **3.7** | **25.6e-2** | **8.1e-2** | **3.2** | 15.4e-2 | 9.0e-2 | 1.7 | 10.3e-2 | 5.4e-2 | 1.9 |
| Pop. dens. | **82.8e-4** | **3.6e-4** | **23.2** | **24.2e-4** | **1.9e-4** | **13.1** | **36.5e-4** | **2.1e-4** | **17.6** | **22.9e-4** | **1.2e-4** | **18.4** |
| Workplace dens. | **35.1e-4** | **2.6e-4** | **13.7** | **18.0e-4** | **1.3e-4** | **13.6** | **12.4e-4** | **1.5e-4** | **8.3** | **49.8e-5** | **8.9e-5** | **5.6** |
| Children | **-3.3e+0** | **1.0e+0** | **-3.2** | **-14.7e-1** | **5.4e-1** | **-2.8** | 3.7e-1 | 6.0e-1 | 0.6 | **-12.2e-1** | **3.6e-1** | **-3.4** |
| Adults | 1.8e+0 | 1.0e+0 | 1.8 | 0.1e+1 | 5.2e-1 | 1.9 | **16.4e-1** | **5.8e-1** | **2.8** | 1.4e-1 | 3.5e-1 | 0.4 |
| Pens | -3.0e-1 | 9.7e-1 | -0.3 | 3.7e-1 | 5.1e-1 | 0.7 | 8.1e-1 | 5.6e-1 | 1.4 | -5.6e-1 | 3.4e-1 | -1.7 |
| Incomes | 2.9e-6 | 3.7e-6 | 0.8 | -1.6e-6 | 1.9e-6 | -0.9 | 1.5e-6 | 2.1e-6 | 0.7 | 0.6e-6 | 1.3e-6 | 0.5 |
| Education | -0.2e+0 | 1.1e+0 | -0.1 | 2.8e-1 | 5.8e-1 | 0.5 | **-17.7e-1** | **6.5e-1** | **-2.7** | 2.8e-1 | 3.9e-1 | 0.7 |
| Unemployment | **4.9e+0** | **1.9e+0** | **2.6** | 3.7e-1 | 9.8e-1 | 0.4 | **2.7e+0** | **1.1e+0** | **2.4** | 7.1e-1 | 6.6e-1 | 1.1 |

Table 7.1: Global regression results for full models. Parameter values significantly different from zero are emphasised.

| | All fires | | D-fires | | E-fires | | N-fires | |
|---|---|---|---|---|---|---|---|---|
| Nr. of observations | 2121 | | 2121 | | 2121 | | 2121 | |
| Bandwidth (in data units) | 1909.7 | | 1757.1 | | 1909.7 | | 710.0 | |
| Effective nr. of parameters | 113.6 | | 129.2 | | 113.6 | | 134.8 | |
| AIC | 11656.0 | | 8886.6 | | 9416.6 | | 7237.8 | |
| Adjusted $r^2$ | 0.51 | | 0.39 | | 0.35 | | 0.35 | |
| Coeff. of determination | 0.53 | | 0.42 | | 0.39 | | 0.39 | |
| Parameter | P-value | Significance | P-value | Significance | P-value | Significance | P-value | Significance |
| Intercept | 0.27000 | | 0.50000 | | 0.15000 | | 0.53000 | |
| Building age | 0.37000 | | 0.68000 | | 0.25000 | | 0.50000 | |
| Building type | 0.00000 | *** | 0.18000 | | 0.15000 | | 0.19000 | |
| Pop. dens. | 0.51000 | | 0.01000 | ** | 0.10000 | | 0.01000 | ** |
| Workplace dens. | 0.03000 | * | 0.25000 | | 0.02000 | * | 0.17000 | |
| Children | 0.00000 | *** | 0.16000 | | 0.12000 | | 0.00000 | *** |
| Adults | 0.02000 | * | 0.00000 | *** | 0.09000 | | 0.10000 | |
| Pens | 0.00000 | *** | 0.00000 | *** | 0.22000 | | 0.00000 | *** |
| Incomes | 0.01000 | ** | 0.06000 | | 0.00000 | *** | 0.05000 | * |
| Education | 0.45000 | | 0.00000 | *** | 0.05000 | * | 0.04000 | * |
| Unemployment | 0.17000 | | 0.02000 | * | 0.70000 | | 0.71000 | |

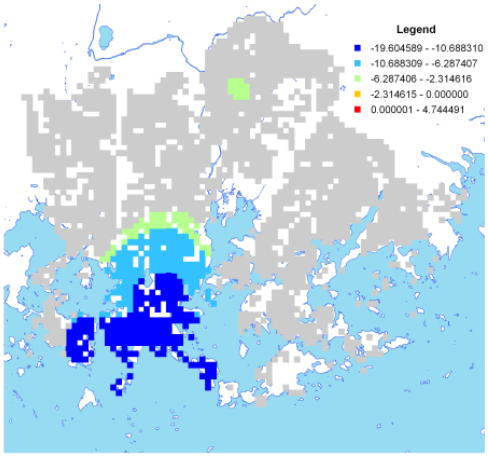
Table 7.2: GWR results for full models. Significance of spatial variability of parameters indicated at 5% (*), 1% (**), 0.1% (***) level.



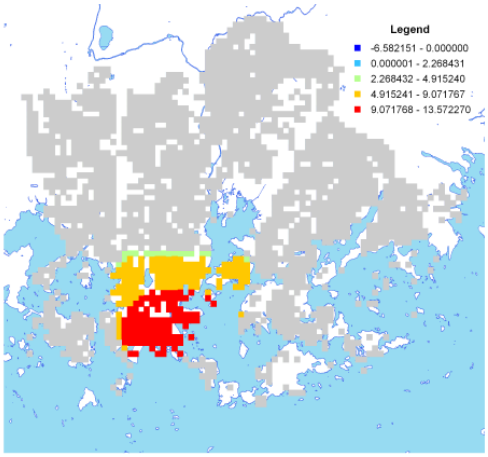

(a) (b)

(c)

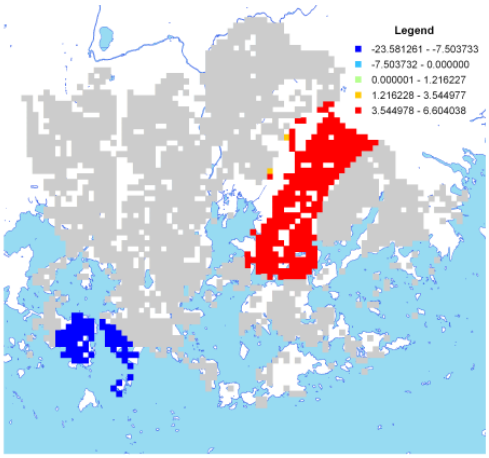
Figure 7.4: Parameter values of GWR for a full model of all fires: (a) Building type; (b) Population density; (c) Workplace density. The blue-green-red colour scale indicates parameter values from negative to positive; the grey masks zero-valued parameters.




(a)

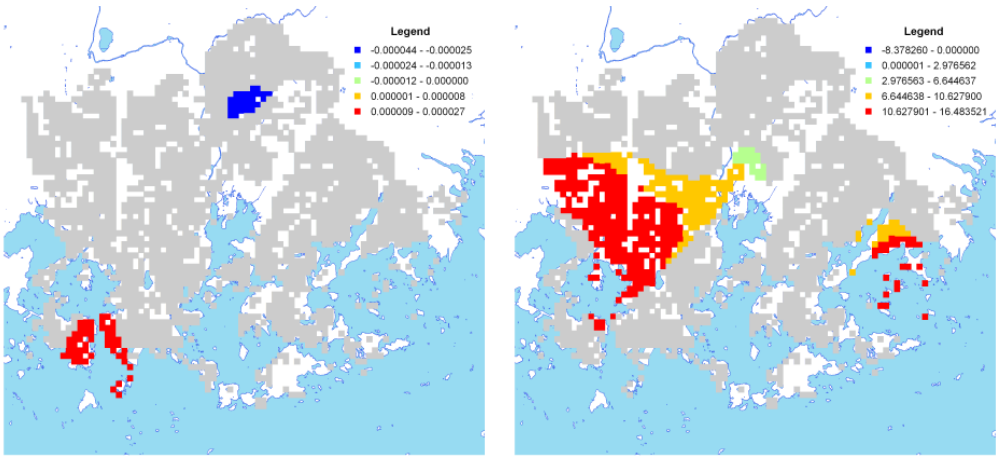



(b)




(c)

Figure 7.5: Parameter values of GWR for a full model of all fires: (a) Children; (b) Adults; (c) Pensioners. The blue-green-red colour scale indicates parameter values from negative to positive; the grey masks zero-valued parameters.




(a) (b)

Figure 7.6: Parameter values of GWR for a full model of all fires: (a) Incomes; (j) Unemployment. The blue-green-red colour scale indicates parameter values from negative to positive; the grey masks zero-valued parameters.

sets separately. The results for the temporal global and GWR models are shown in Tables 7.1 and 7.2, respectively. The following is observed. In all cases, the GWR model significantly reduces the AIC values. However, a considerable deal of variation in the density of fires is not accounted for by the models. As the parameter values and their significance vary on the time axis, indicating temporal changes in the influence of predictor variables on the distribution of domestic fires, the temporal models are further analysed separately.

Excluding one variable step by step while comparing AIC values enables the complexity of the models to be reduced and the best-fitting model to be found. The results are listed in Tables 7.3 and 7.4. Compared to the full models, the reduced GWR models show a decrease in the AIC values with 184.2 for the d-fires, 187.9 for the e-fires and 195.1 for the n-fires. In this way, the population density, density of workplaces, and household structure are identified as significant predictor variables. The $t$ values confirm the significance of the variables. The coefficients of determination of the reduced GWR models show a small increase in the variation in the density of the fires accounted for in the models, with 6% for d-fires, 5% for e-fires, and 4% for n-fires. The reduced models therefore capture the relationships in the data better

than the full models. Apart from the density of workplaces in the model for e-fires, the predictor variables exhibit significant spatial variation; see Table 7.4.

The output of GWR models for the selected temporal categories is mapped in Figures 7.7–7.12. GWR models provide better fit than global models in the central part of the study area and in the north for d-fires, east for e-fires, and east and west for n-fires. Standardised residuals exhibit no distinct autocorrelation in all cases. The parameter values for population density are significantly different from zero for the largest part of the study area. The values are positive; they are highest in the north for d-fires, in the centre and north-east for e-fires, and in the centre and north-west for n-fires. The influence of the density of workplaces is also positive; it is most significant in the north, with only small temporal variations. The parameters representing the influence of households with children are negative in the central part of the study area. Additional areas with positive values are found in the west and in the east for e- and n-fires. Households with adults represent a significant parameter only in the case of d-fires, with positive values in the central part of Helsinki.

This study demonstrates that the relationships between the distribution of domestic fires and background phenomena change with both location and time. This is an important message for the fire & rescue services. The investigation of these spatio-temporal variations brings new viewpoints into the relationships between the phenomena and enhances their understanding. In this way, the knowledge of the background processes being studied, which is necessary for the development of a realistic risk model, can be supported.

## 7.4 Conclusions

### 7.4.1 Capturing spatial and temporal aspects

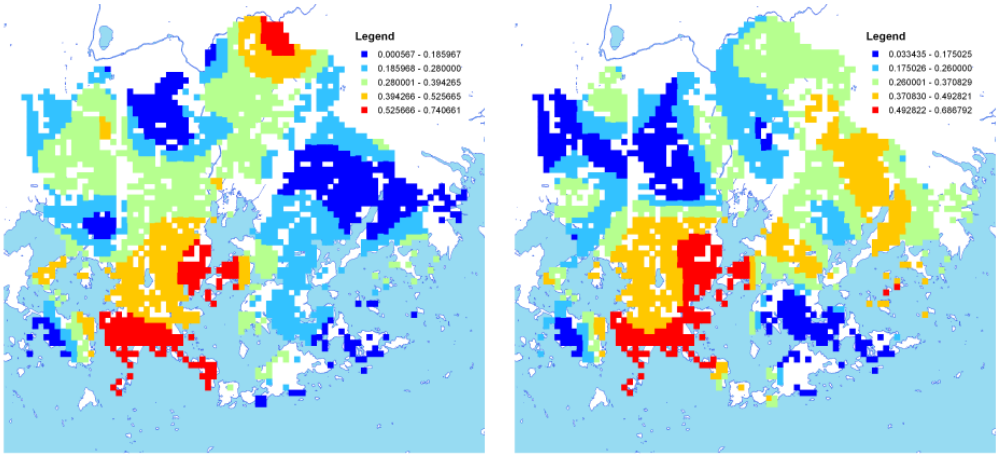
GWR is found to be useful for analysing and modelling any spatial processes, as it is truly a spatial technique, which also allows spatial variations in the relationships to be viewed. The model diagnostics provided by GWR and the possibility of mapping the results enable interesting locations to be identified for further investigation. The question is whether the observed variations are due to locally different spatial behaviour, or simply due to model misspecification. In either case, an examination of the nature of spatial variation can suggest a more accurate model specification and

| | D-fires | | | E-fires | | | N-fires | | |
|---|---|---|---|---|---|---|---|---|---|
| AIC | 9075.8 | | | 9573.3 | | | 7369.8 | | |
| Adjusted $r^2$ | 0.28 | | | 0.26 | | | 0.26 | | |
| Coef. of determination | 0.29 | | | 0.26 | | | 0.26 | | |
| Parameter | Estimate | St Err | T | Estimate | St Err | T | Estimate | St Err | T |
| Intercept | 41.8e-2 | 9.6e-2 | 4.3 | 58.8e-2 | 9.2e-2 | 6.3 | 24.8e-2 | 5.5e-2 | 4.5 |
| Pop. dens. | 23.6e-4 | 1.8e-4 | 13.3 | 38.7e-4 | 1.7e-4 | 22.8 | 23.6e-4 | 1.0e-4 | 23.4 |
| Workplace dens. | 19.2e-4 | 1.3e-4 | 15.3 | 13.0e-4 | 1.4e-4 | 9.3 | 58.6e-5 | 8.3e-5 | 7.0 |
| Children | -17.4e-1 | 2.7e-1 | -6.4 | -7.8e-1 | 3.0e-1 | -2.6 | -9.6e-1 | 1.8e-1 | -5.5 |
| Adults | 10.2e-1 | 2.2e-1 | 4.6 | | | | | | |

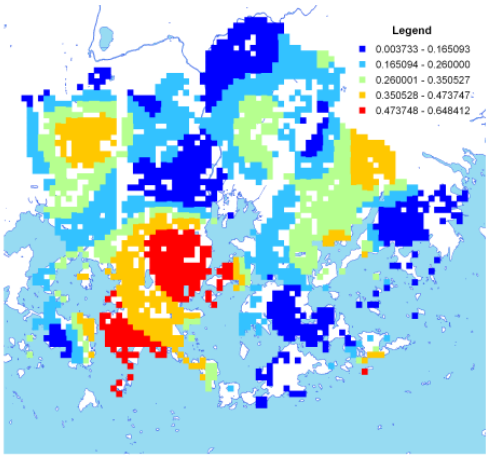
Table 7.3: Global regression results for reduced models.

| | D-fires | | E-fires | | N-fires | |
|---|---|---|---|---|---|---|
| Nr. of observations | 2121 | | 2121 | | 2121 | |
| Bandwidth (in data units) | 1016.6 | | 1016.6 | | 1016.6 | |
| Effective nr. of parameters | 149.7 | | 119.6 | | 119.6 | |
| AIC | 8702.4 | | 9228.7 | | 7042.7 | |
| Adjusted $r^2$ | 0.44 | | 0.41 | | 0.40 | |
| Coef. of determination | 0.48 | | 0.44 | | 0.43 | |
| Parameter | P-value | Significance | P-value | Significance | P-value | Significance |
| Intercept | 0.00000 | *** | 0.01000 | ** | 0.02000 | * |
| Pop. dens. | 0.01000 | ** | 0.01000 | ** | 0.01000 | ** |
| Workplace dens. | 0.00000 | *** | 0.10000 | | 0.01000 | ** |
| Children | 0.00000 | *** | 0.00000 | *** | 0.00000 | *** |
| Adults | 0.01000 | ** | | | | |

Table 7.4: GWR results for reduced models. Significance of spatial variability of parameters indicated at 5% (*), 1% (**), 0.1% (***) level.

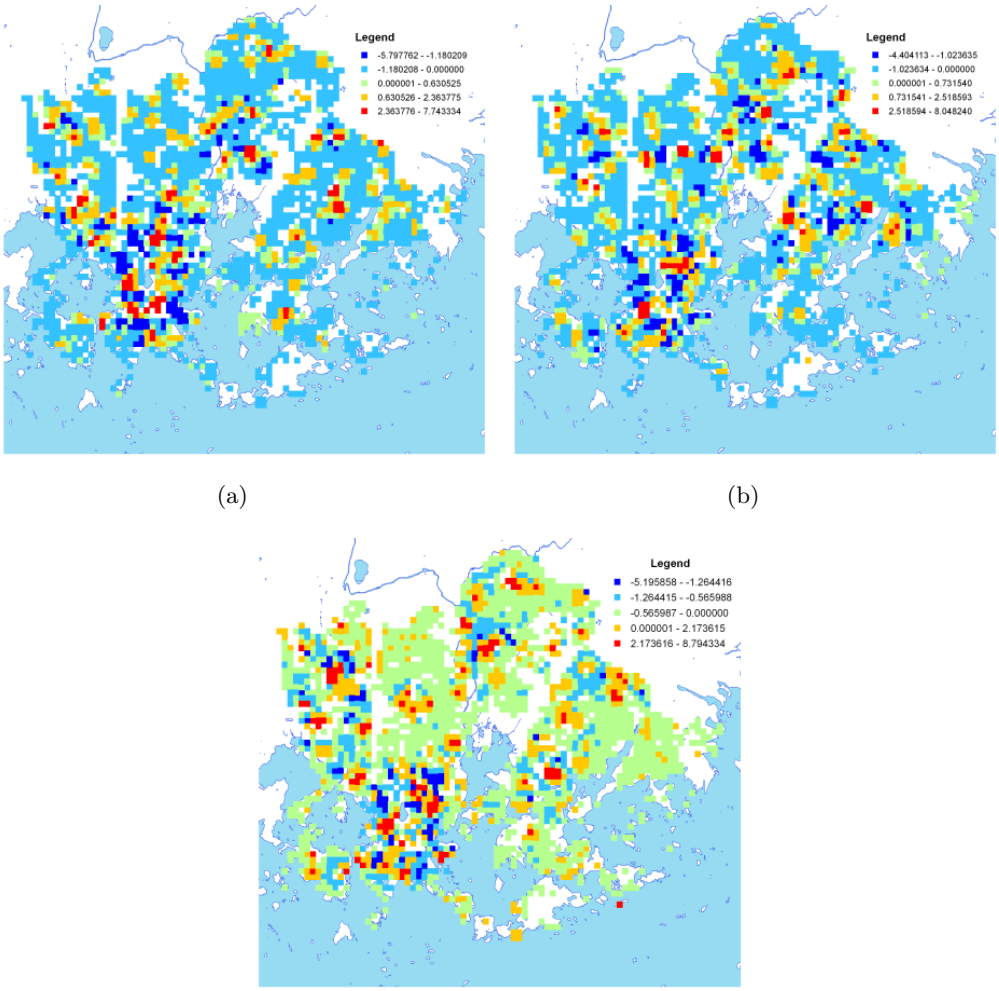



(a) (b)




(c)

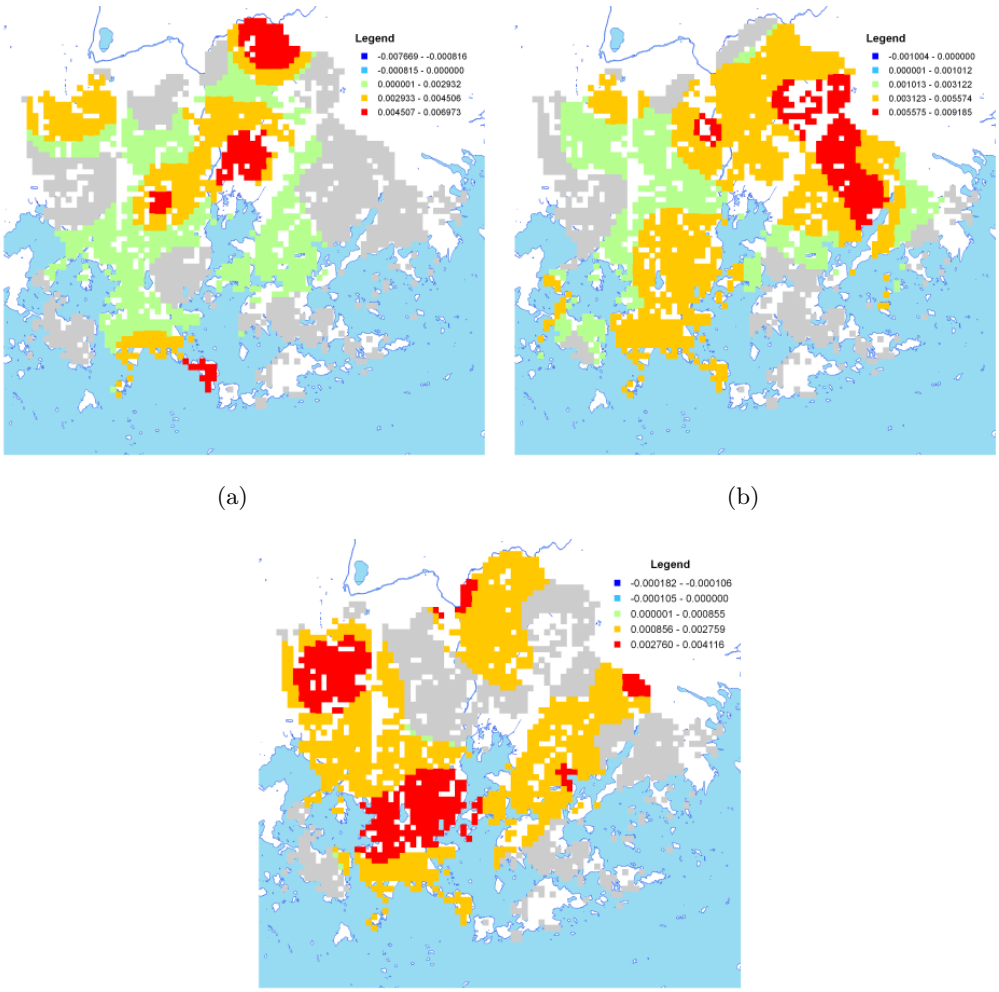
Figure 7.7: Local $r^2$ for reduced models: (a) d-fires; (b) e-fires; (c) n-fires. The red-green-blue colour scale indicates where the GWR provides a better fit compared to the global models (red-green areas).




(a) (b)

(c)

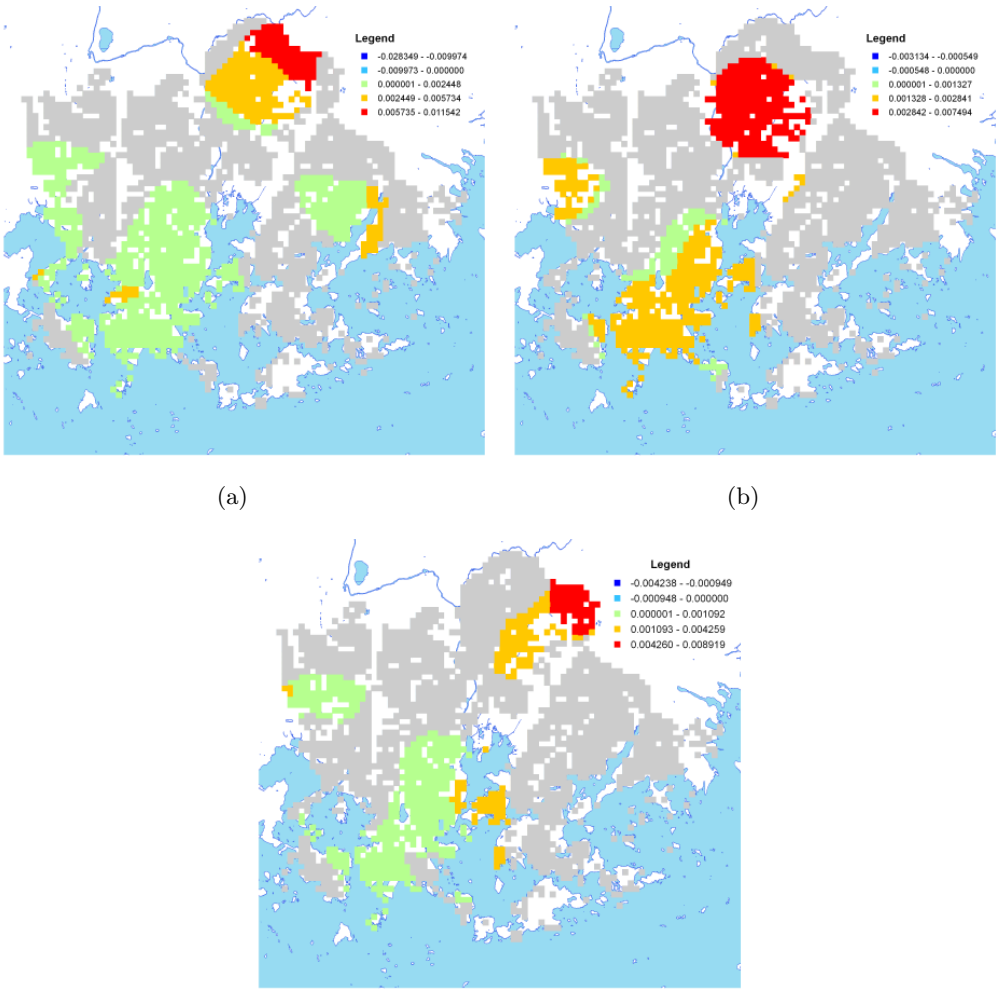
Figure 7.8: Standardised residuals for reduced models: (a) d-fires; (b) e-fires; (c) n-fires.




(a) (b)

(c)

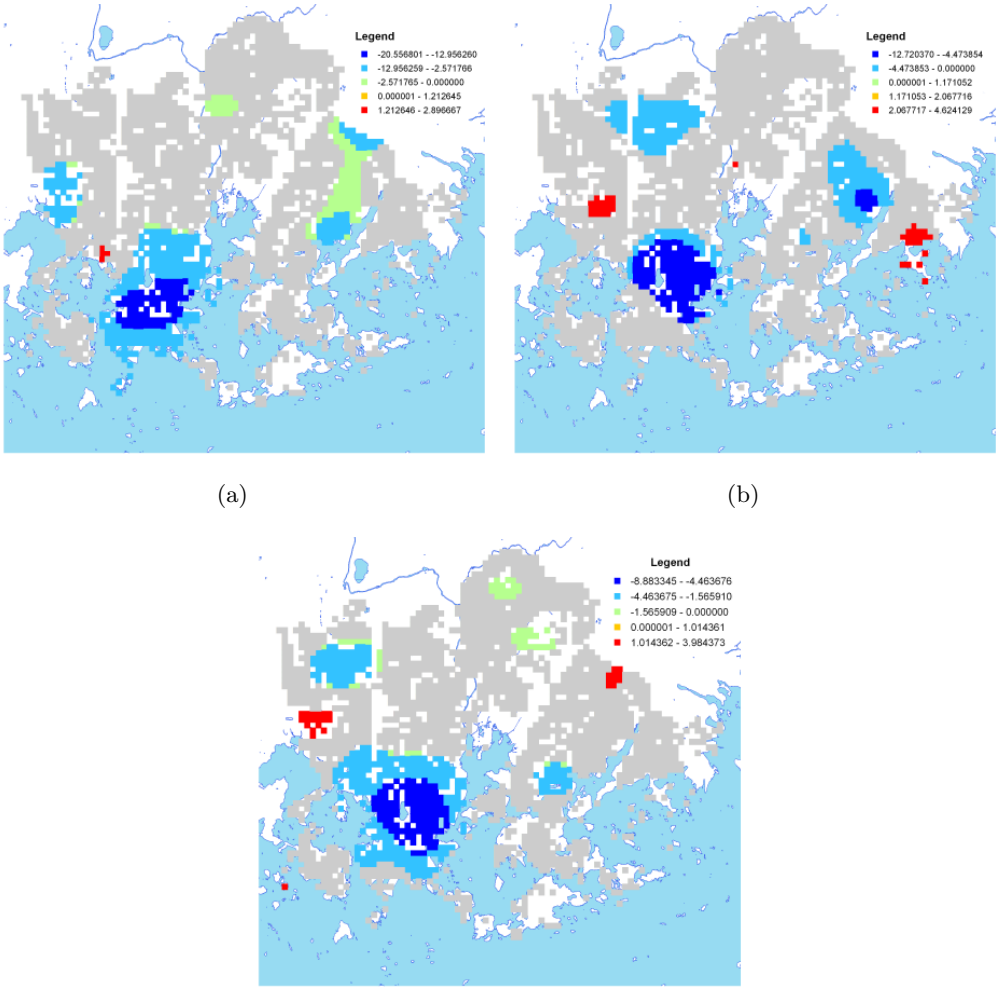
Figure 7.9: Parameter values of population density for reduced models: (a) d-fires; (b) e-fires; (c) n-fires. The red-green-blue colour scale indicates positive-negative parameter values, while grey masks zero-valued parameters.




(a) (b)

(c)

Figure 7.10: Parameter values of density of workplaces for reduced models: (a) d-fires; (b) e-fires; (c) n-fires. The red-green-blue colour scale indicates positive-negative parameter values, while grey masks zero-valued parameters.




(a) (b)

(c)

Figure 7.11: Parameter values of households with children for reduced models: (a) d-fires; (b) e-fires; (c) n-fires. The red-green-blue colour scale indicates positive-negative parameter values, while grey masks zero-valued parameters.

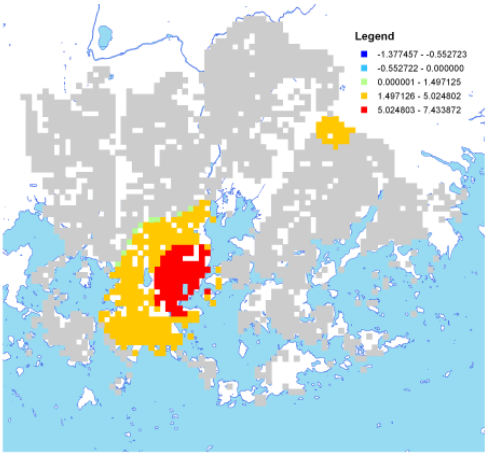



Figure 7.12: Parameter values of households with adults for a reduced model of d-fires. The red-green-blue colour scale indicates positive-negative parameter values, while grey masks zero-valued parameters.

improve our knowledge of the system under investigation [Fotheringham et al., 2002].

Capturing the temporal aspect with GWR is feasible by splitting the dataset into pre-defined categories, which are modelled separately. Such an approach requires prior knowledge about suitable time categories. The results are explored in order also to compare variations on the temporal scale. Using more sophisticated methods for analysing the spatio-temporal results of GWR, e.g. [Demšar et al., 2008] is advisable.

### 7.4.2 Type of knowledge discovered

GWR is a local multivariate statistical method for analysing large and complex spatial data. As a version of regression analysis, GWR provides an insight into the dependence relationship between predictor and response variables, while also quantifying the relationship. In contrast to global methods, GWR enables local spatial variations in the process to be detected.

As GWR enables the non-stationarity of the processes to be modelled by allowing the regression parameters to vary spatially, it offers a more easily interpretable solution to the problem of spatial autocorrelation. It also provides local estimates of spatial autocorrelation directly. In addition, GWR, being thought as a 'spatial microscope', enables previously unimagined details to come into focus, while the local

models appear to be more robust regarding scale changes. In this way, GWR comprises the crucial issues of spatial non-stationarity, spatial dependence, and spatial scale [Fotheringham et al., 2002].

GWR models are based on the statistical testing of parameter significance, which represents solid ground. A number of associated diagnostics enables various aspects of the process being modelled to be explored, which supports its understanding. In this way GWR enables important variables characterising any spatial processes to be discovered and also quantified.

### 7.4.3 Weaknesses

GWR, as a local modelling technique, provides a specific insight into the processes being investigated. In contrast to averaged global analysis, it allows local variations in relationships to be measured. This creates new demands for the user as concerns the correct interpretation of the GWR results and understanding of the processes being modelled.

### 7.4.4 Requirements for the user

The application of GWR requires an insight into spatial statistics on the part of the user. On the other hand, with experience of regression analysis, it is relatively easy to become familiar with the core idea of GWR. The suitability of the GWR results for being mapped makes the data exploration intuitive, which leads to the construction of appropriate spatial models.

GWR is supported by user-friendly software, which can be run under Windows and Unix, and a code for GWR in R is available from the authors. A GWR tool is also available in the Spatial Statistics Toolbox of ArcGIS 9.3. Because of its support for a file interchange format, the output of GWR can easily be mapped using common GIS.

Conceptualisation and data pre-processing are necessary to some extent before the application of GWR. In addition, suitable visualisation and the proper interpretation of the results require expert skills. A GIS professional therefore plays a role in this application too.

# Chapter 8

# Self-Organising Maps

This chapter is based on the following article. The author contributed to the analysis design and played a leading role in the data pre-processing, the performance of the analysis, and the writing of the paper.

Špatenková, O., Demšar, U., Krisp, J. (2007) Self-Organising Maps for exploration of spatio-temporal emergency response data, Proceedings of Geocomputation 2007, Maynooth, Ireland, 2007.

## 8.1 Method

The Self-Organising Map (SOM) is an artificial neural network, which is capable of distinguishing similarity patterns in multidimensional space. It defines a mapping of multidimensional data onto a lattice with a low number of dimensions, usually two, while preserving both probability distribution and the topology of the input data [Kohonen, 1997; Silipo, 2003]. As the SOM offers an alternative view of multidimensional data, it is a useful tool for knowledge discovery. It has recently been used in a number of geographical applications, e.g. [Jiang and Harrie, 2004; Koua and Kraak, 2004; Guo et al. 2005; Demšar, 2007].

The SOM consists of network nodes – neurons, which are usually arranged in a rectangular or hexagonal lattice. Each neuron represents a model, or a prototype of the input data computed by the SOM algorithm during the training phase. Next, in the mapping phase, SOM automatically classifies a new input data vector to a single

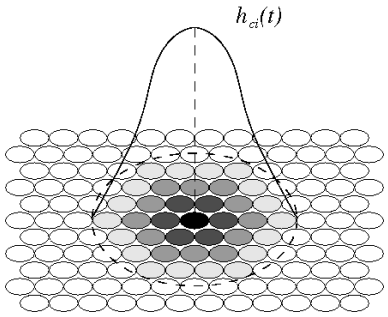



Figure 8.1: Neighbourhood function $h_{ci}(t)$ centred on the best matching neuron (coloured in black) in a SOM organised in a hexagonal lattice.

winning neuron.

The SOM uses unsupervised learning to arrange the neurons. In contrast to supervised learning, where the pairs of input data and desired output values are used for training, this is a purely data-driven approach. At the beginning, the initial vectors of weights are assigned to the neurons either randomly or using principal component analysis [Jolliffe, 2002]. The latter approach is used to speed up the training, as the initial weights provide a good approximation of the final values. The following process, based on competitive and cooperative learning, is repeated in a large number of iterations: the SOM finds the best matching neuron for each input and recalculates the vectors of weights $\boldsymbol{w_i}(t)$ in its neighbourhood according to the formula

$$\boldsymbol{w_i}(t+1) = \boldsymbol{w_i}(t) + h_{ci}(t)\alpha(t)(\boldsymbol{X}(t) - \boldsymbol{w_i}(t)), \tag{8.1}$$

where $\boldsymbol{X}(t)$ is the input vector, $\alpha(t)$ is a monotonically decreasing learning coefficient, and $h_{ci}(t)$ is the neighbourhood function, which decreases towards 0 with the distance from the neuron $\boldsymbol{w_i}$ to the best matching neuron $\boldsymbol{w_c}$ (Figure 8.1). The range of the neighbourhood function shrinks with each iteration step $t$, so that the weights of the neurons are converging to the local estimates at the end of the training process. The initial neighbourhood is usually fairly wide in order to facilitate a global ordering. It shrinks with a monotonically decreasing function during the first phase, when the proper ordering takes place. During the fine adjustment phase, the neighbourhood can still contain the nearest neurons of the neuron $\boldsymbol{w_c}$. In this way nearby cells activate each other up to a certain distance to learn from the same input, which results in similar input data being mapped to neighbouring neurons [Kohonen, 1997].

Since the SOM projection reduces the dimensionality of the original dataset, information is lost during the process. As the mapping balances between the accuracy of data representation and the topological accuracy, two evaluation criteria are used to assess the quality of the SOM. Different approaches to measure them exist, usually based on the training data given. The average quantisation error and the topographic error are chosen for their simplicity. The former, representing the average distance between the data vectors and their best matching neurons, measures the map resolution. The latter denotes the percentage of data vectors, for which the first- and second-best-matching neurons are not adjacent, and thus measures the preservation of the topology [Kiviluoto, 1996].

The two-dimensional output space of the SOM offers a wide range of possibilities for visual inspection, as described in [Vesanto, 1999]. A distance matrix is used in this study to provide an overview of the SOM structure. It represents the differences between the immediate neighbours in the SOM and thus enables clusters in the data to be found. The differences are indicated in the distance matrix by a colour or a grey shade with similar cells (clusters) shown as light areas, and dissimilar cells are dark, denoting boundaries between the clusters.

Data histograms representing the best matching neuron for each data sample provide an insight into the distribution of the dataset on the SOM. There are several ways to visualise data histograms, e.g. as a 3-D bar graph with the height of the bar corresponding to the value of the data histogram or as a dot plot, where each data sample is plotted on top of the corresponding neuron with a small random offset to make the samples distinguishable. Another suitable visualisation of data histograms, which also avoids possible problems with overprinting, uses symbols with sizes proportional to the number of hits in the corresponding neuron.

The properties of the SOM can also be described by the attribute values of all the neurons in the SOM. Component planes, one for each attribute, play the key role here. They can be visualised using a colour scale to represent the changes in the attribute values. As the SOM consists of prototypes characterising the data, the component planes show the distribution of the data values in the SOM.

Specific objects are linked between the individual visualisations of SOM according to their position. By linking the component planes to the distance matrix, cluster properties can be identified. The component planes can be further used for finding correlations between the attributes by matching similar patterns in identical positions

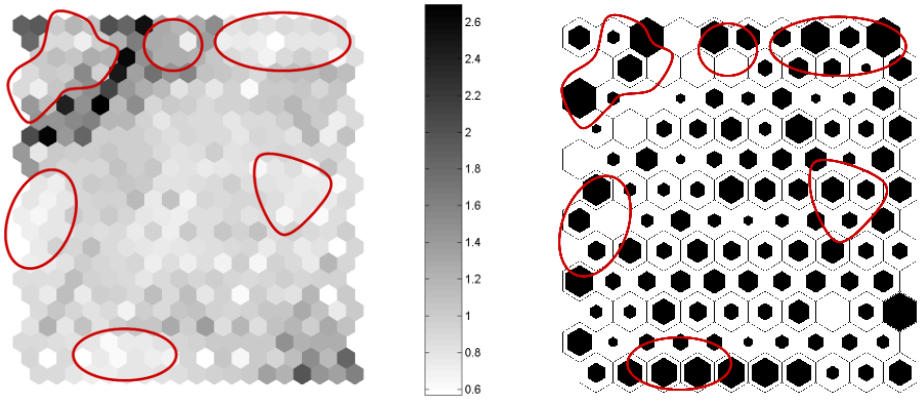
of the component planes. As this approach is rather vague, interesting component combinations can be selected for detailed investigation using scatter plots [Vesanto, 1999; Koua and Kraak, 2004].

The SOM toolbox for Matlab was used for the SOM calculation and the visualisation of the results [Vesanto et al., 2000].

## 8.2 Data pre-processing

Two approaches to the analysis are used to demonstrate the influence of a different conceptualisation to the results. First, the SOM is used to visualise the incident dataset in order to identify clusters in the distribution of domestic fires. Each incident is, in addition to X- and Y-coordinates, characterised by the time of its occurrence. The temporal analysis focuses on the hourly, day-of-the-week, and monthly scales. The incident records are joined to the additional attributes based on the location. The input data vector of the SOM represents the incident record with the associated attributes of the nearest building (type and age) and socio-economic attributes of the corresponding census cell (population density, workplace density, ratio of households with children, adults, and pensioners, average incomes, education of the inhabitants, and unemployment rate). This approach allows the characteristics of the domestic fires to be studied.

The risk analysis, however, concerns variations in the distribution of incidents all over the study area. Therefore, in the next approach, the study area is divided into grid cells, which constitute the input of the SOM. Census records determine the form of the grid, in which building attributes are associated with the average age and prevailing building type within the grid cell. The distribution of domestic fires is represented as a kernel density surface projected onto the grid. The hourly scale is chosen to add a temporal dimension to the analysis. Additional attributes therefore represent the density of n-fires (1 a.m. – 8 a.m.), d-fires (8 a.m. – 5 p.m.), and e-fires (5 p.m. – 1 a.m.). Such a conceptualisation allows correlations to be found between the spatio-temporal distribution of domestic fires and the background environment.




(a) (b)

Figure 8.2: Clusters identified from the distance matrix as light areas delimited by darker colours (a) and corresponding data histogram (b) of the SOM for the incident dataset.

## 8.3 Results

Similarities existing in the original spatio-temporal space form clusters, which can be identified by the SOM and visualised as light areas in the distance matrix. Figure 8.2(a) shows the distance matrix of the SOM applied to the incident records with identified clusters marked in red. Figure 8.2(b) shows the corresponding data histogram, where the marker size indicates the number of hits in each cell. It can be seen that clusters identified from the distance matrix represent concentrations of incidents that are similar to each other, while empty cells correspond to the boundaries between clusters.

In order to characterise the clusters that are identified, these are superimposed on the component planes, which represent attribute values of the prototype neurons of the SOM (Figure 8.3). It can be observed that the clusters match the patterns in the component planes to some extent. For example, the most significant cluster in the top left-hand corner represents domestic fires that happen in areas with a low population density and a low to medium density of workplaces. This cluster concerns new buildings which are classified as being for work or leisure. The incidents in this cluster occur during the daytime, on working days, and mostly in the winter or spring months. They are located in the southern part of the study area, but are spread all over it in the east-west direction.

The bottom-most cluster corresponds to incidents occurring in highly populated areas in housing buildings. These fires are located in the south-west and happen during the daytime. The households in question are occupied mainly by adults with high education but low incomes. This cluster can also be characterised by a medium unemployment rate.

The middle cluster on the left-hand side represents fires in areas with a high density of workplaces. These are located in the south-west. Naturally, the fires concern work-type middle-aged buildings. On the temporal scales the fires occur on working days during the daytime.

The middle cluster at the top embraces fires in buildings classified as being for work and which were constructed recently. They are located in the north-eastern part of the study area. The density of workplaces, however, is low there. This cluster can be further characterised by a high ratio of households with pensioners.

The two remaining clusters on the right-hand side of the SOM represent fires in housing buildings in areas with low to medium population densities. The fires in both clusters occur in the evening hours. The upper cluster is located more to the north and concerns newer buildings than the lower one. The clusters also differ in the socio-economic structure of their inhabitants, which is more favourable in case of the top cluster.

Comparing the component planes can reveal correlations between the attributes. For example, as shown in Figure 8.4, while daytime fires (represented as green areas) occur in the southern part of the study area, evening fires (represented as red areas) are located in the north. Further, evening fires concern mostly housing buildings which were constructed recently in areas with a low density of workplaces. Daytime fires, in contrast, occur in areas with a high density of workplaces, in all types of buildings. Daytime fires are also related to areas with a low density of households with children or pensioners, but a high density of adults.

To get an insight into the relations between the distribution of incidents within the background environment, the SOM is also applied to the grid cells representing the attributes being studied in particular locations in the study area. Figure 8.5 shows the distance matrix with identified clusters and the corresponding data histogram.

The location of the identified clusters in the component planes is shown in Figure 8.6. Most of the clusters in the data, however, cover areas with a low density of fires. Only the cluster in the bottom left represents an area with a high density

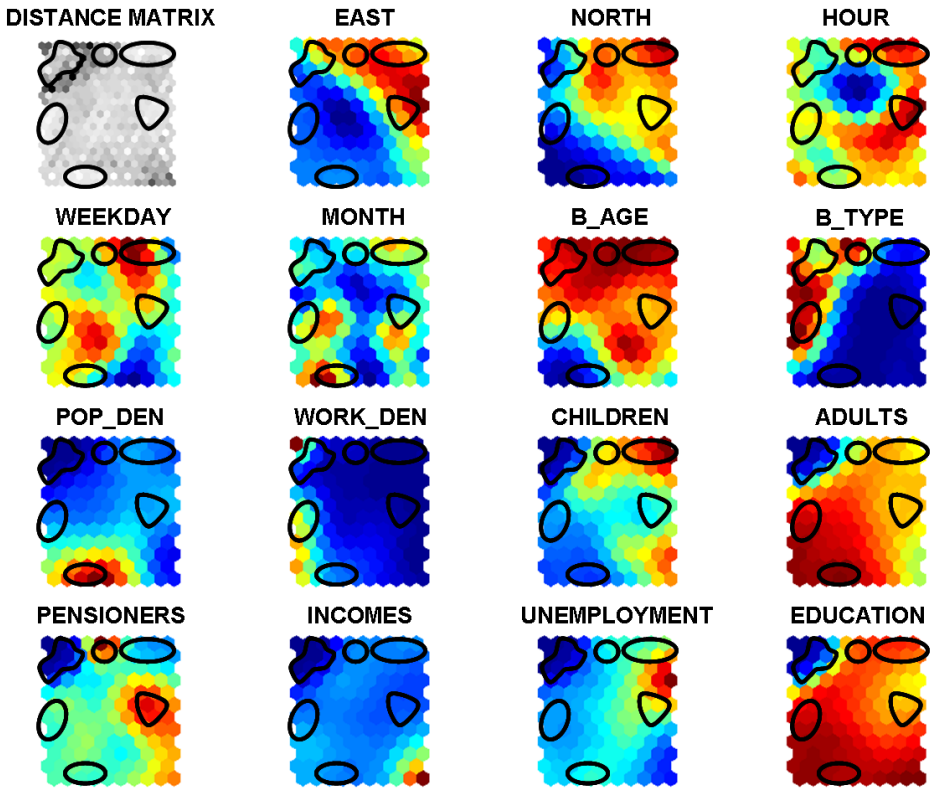



Figure 8.3: Location of the identified clusters in the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

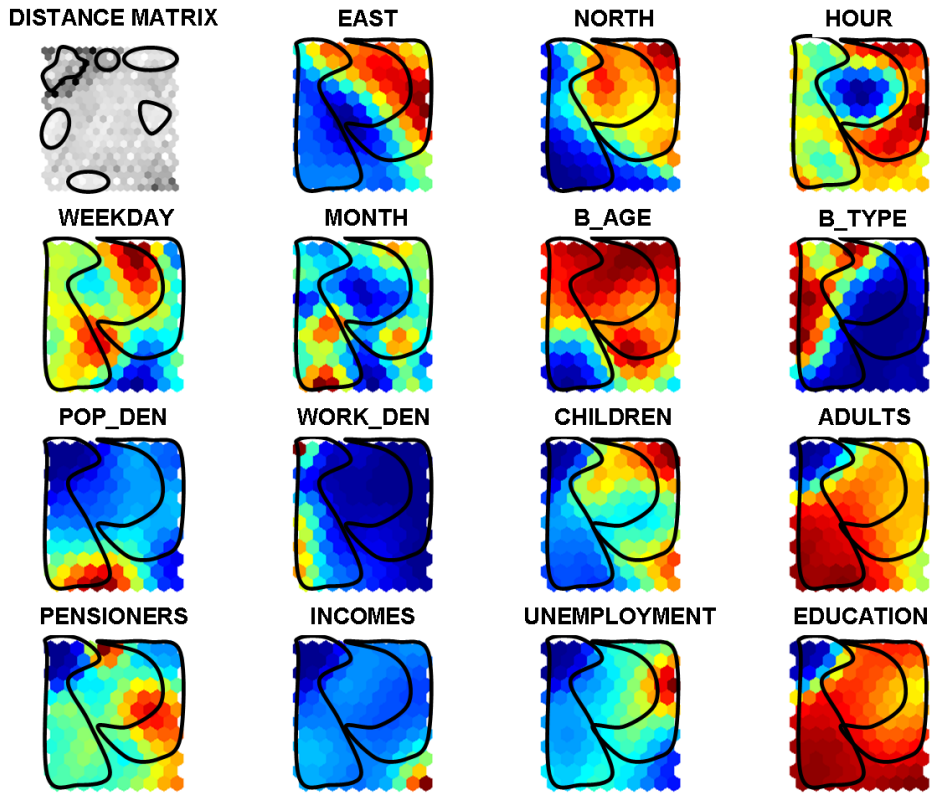
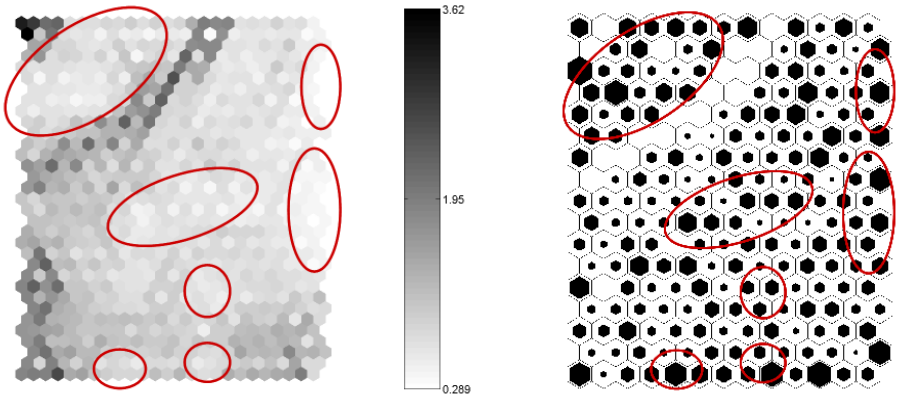



Figure 8.4: Discovering relationships between the attributes from the component planes. The green and red colours in the HOUR component plane indicate morning and evening fires, respectively. The characteristics of these incidents can be found from the remaining component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).




(a) (b)

Figure 8.5: Clusters identified from the distance matrix (a) and corresponding data histogram (b) of the SOM for the grid representation of the data.

of e-fires and a medium density of d-fires, which is located in the central part of the study area. Fires in this cluster occur in housing buildings in areas with average population densities. The households falling into this cluster are inhabited mainly by adults with low incomes, average education, and a medium unemployment rate. The density of fires is low at night.

A study of the component planes indicates striking differences between the temporal categories of domestic fires under study (Figure 8.7). In contrast to d-fires, which occur mainly in the south-western part, e-fires are spread over the study area. E-fires also affect mostly residential areas, while d-fires happen in all building types. Population density and poorer socio-economic conditions play a role in the occurrence of e-fires, whereas d-fires are related rather to the density of workplaces. There is a strong positive correlation between n-fires and population density. In addition, n-fires frequently occur in housing buildings in the east-central part of the study area. This area is scantily populated and exhibits positive socio-economic conditions. Adult households seem to contribute to the occurrence of fires; however, there are also areas with a high density of adults where the density of fires is low.

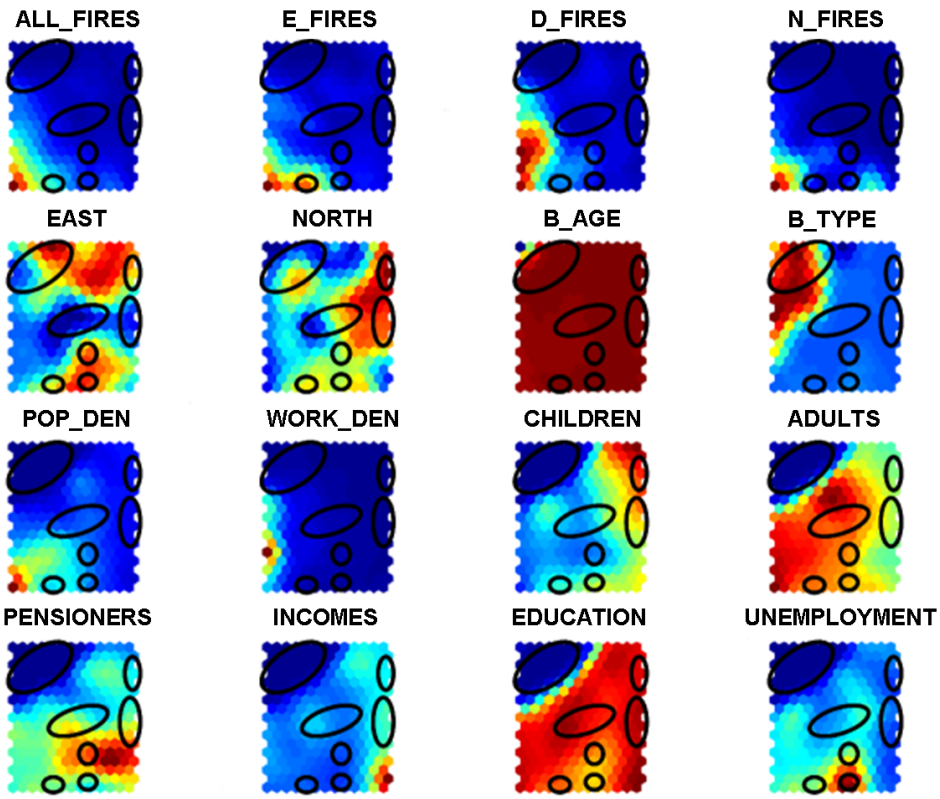



Figure 8.6: Location of the identified clusters for the grid representation of the data in the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

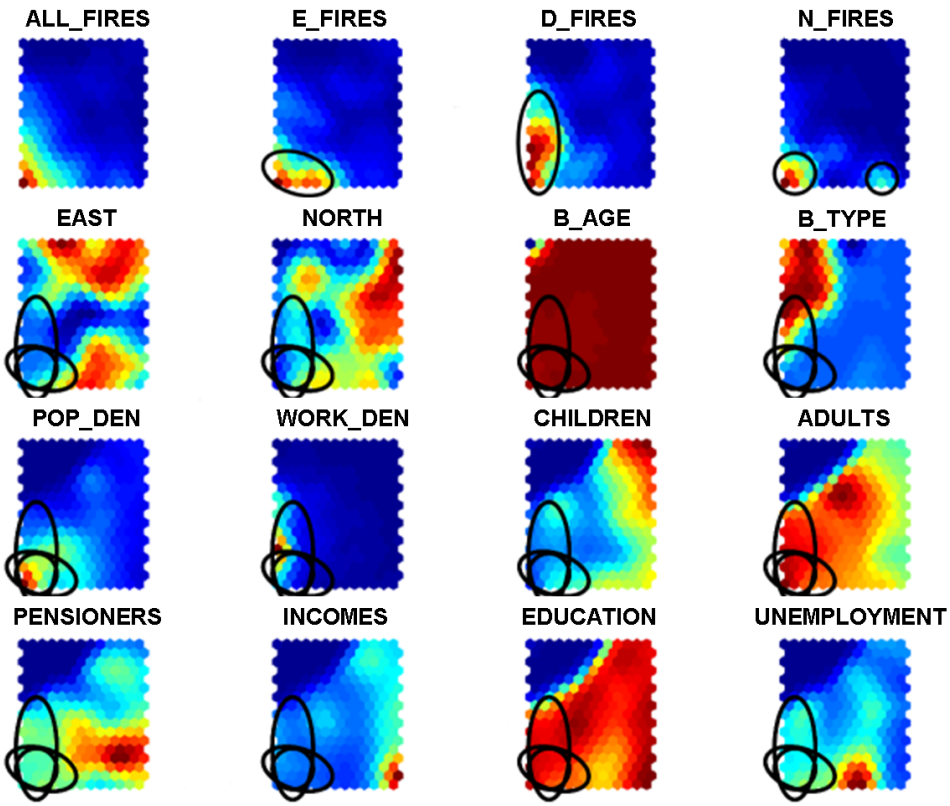



Figure 8.7: Identifying differences between e-fires, d-fires, and n-fires in the grid representation from the component planes. The colour scale goes from dark blue (low values) through green (medium values) to dark red (high values).

## 8.4 Conclusions

### 8.4.1 Capturing spatial and temporal aspects

The SOM is designed to visualise complex multidimensional data in a two-dimensional space. Because of its topology-preserving property, it is suitable for analysing geographical data. This chapter demonstrates two conceptually distinct approaches to how the SOM can be applied. The first approach focuses on the incident data with associated background information with the aim of characterising the dataset. As the original point dataset is used, the method takes advantage of the level of detail offered by the actual data. The temporal aspect is integrated into the analysis on several scales.

The other approach aims to analyse the distribution of the phenomenon being studied within the background environment. In contrast to the previous approach, it also considers those parts of the study area where no incidents occurred. The spatial aspect is represented in this case by grid cells of suitable resolution covering the study area. The variables that are studied are represented as density surfaces to avoid problems related to rasterisation and attached as attributed to particular grid cells. Time is incorporated on predefined scales as additional attributes.

### 8.4.2 Type of knowledge discovered

The nub of the SOM is a computational clustering algorithm which offers a wide range of possibilities to present the results visually. In this way it provides a fast qualitative insight into the patterns existing in the data.

Mapping using the SOM is multidimensional, as it treats all the attributes at the same time. The SOM yields a self-categorisation of the input dataset on the basis of identified similarities. Besides the probability density, clustering using the SOM also preserves topological relations, so that similar patterns are mapped onto the nearby cells. The SOM also enables correlations to be found between the attributes being studied by comparing the component planes.

### 8.4.3 Weaknesses

The SOM is often seen as a black box by potential users, which may make its application difficult. It is therefore important to understand the principles of the algorithm

behind it in order to ensure that the SOM meets the user's expectations.

In general, the SOM can be interpreted as an approximation of the input space so that similar data items are located in the nearby cells of the SOM. There are many ways to display the SOM properties visually that enable clusters and local relationships in the data to be found. On the other hand, the visual exploration enables only the strength of the relationships discovered to be estimated.

The SOM can be regarded as a mapping of multidimensional space onto a two-dimensional lattice, which necessarily results in information loss. The individual cells of the SOM are prototypes characterising the data, which generally represent more data items. This prevents the details from being observed. For example, it is difficult to distinguish the particular values of coordinates, and temporal attributes can only be estimated in general terms such as daytime, evening, or night. Such patterns are eventually to be investigated further.

The visualisations of the SOM can be seen as a spatialisation of the original dataset, although, they have nothing to do with the original geographic space and a connection of the SOM to the real-world locations can be difficult to infer. The software used in this study does not support the integration of the SOM visualisations with the original map, which would make the interpretation easier.

### 8.4.4 Requirements for the user

As illustrated in this study, the conceptual design of the analysis is essential in the application of the SOM to geographical data. The aim of the inspection must first be determined and then followed by appropriate data pre-processing. A GIS expert plays an indispensable role in this process.

As in the case of any application of advanced analysis methods, the use of the SOM requires a basic knowledge of the algorithm being used. In a way, the SOM behaves like a black box, but the user must understand what to expect from the results. Some skills are also needed for proper interpretation yet the SOM facilitates a visual representation of the results which makes the exploration easier.

The SOM toolbox for Matlab used in this study is available free of charge under the GNU General Public Licence. It was developed in particular for data mining purposes, and is supported by various powerful visualisation functions. However, the SOM toolbox does not provide any connection to the original map to ease the interpretation

of spatial data. There exist software packages that integrate the SOM with other geovisualisations, e.g. GeoVista Studio [Gahegan et al., 2002]. Such applications suffer, though, from non-flexibility in the setting of the SOM algorithm.

# Chapter 9

# Discussion

This study presents a collection of methods to unveil relationships from geospatial data. The aim is to cover a wide spectrum of methods, each of which originates in a different scientific domain. The first method presented, the visual mining of geospatial data, relates to the developments in the field of information visualisation. The following methods, in contrast, represent quantitative approaches to the representation of the patterns in the data. The analysis of contingency tables is a traditional statistical method, while point pattern analysis and GWR are developed with respect to the special nature of geospatial data. Finally the SOM, originating in the field of artificial intelligence, represents a combination of computational and visual approaches to data mining. Clearly, this list of suitable methods is not exhaustive. Other methods that may provide a good solution to the given problem exist, e.g. association rules [Karasová et al., 2005] or logistic regression [Andrews et al., 2003; Preisler et al., 2004].

Similarly, the study does not aim to cover the entire data mining framework, which includes many other tasks and techniques. As data mining is heavily application-dependent, domestic fires are used as a case study in this research to demonstrate how the methods selected support the formation of the knowledge needed for the risk modelling. The research could therefore continue by analysing and comparing other techniques to explore what new aspects they may bring to the knowledge being discovered.

The main focus of the study, however, is on the higher-level concepts. The new insights each of the methods brings to the relationships between the data are the focus

of interest, while their application to domestic fires only illustrates the research. The study can therefore be used for formulating hypotheses about the patterns observed from the data. Consistent with the principles of the knowledge discovery process, the verification of the hypotheses in cooperation with domain experts is necessary before any action is taken. This step requires further research, which is beyond the scope of the current study.

The case study, covering the city of Helsinki, is used to demonstrate the use of the methods. Although of particular interest to the fire & rescue services, Helsinki is clearly distinct from other regions in Finland. The results expressing the relationships between domestic fires and various underlying aspects identified in this study cannot be generalised. However, the methods are also applicable in other cities or regions where a sufficient number of past incident records are available. Their peculiar geographical and socio-economic characteristics, which may be different from those of Helsinki, must be taken into consideration.

The methods presented in this study are general and could be used in further applications for which probabilistic approaches can be used. After a suitable conceptualisation their application can be extended to other phenomena, such as crime distribution, house prices, the occurrence of disease, or animal behaviour.

Knowledge discovery in the realm of geospatial data brings new challenges to the process, such as geographical and temporal data compatibility, including semantics, precision or geometry, the representation of complex spatial and temporal objects and relationships, scalability, or the availability of suitable and user-friendly tools. As shown in this study, there is a need for a proper geospatial conceptualisation before the data mining techniques can be applied. Thus, in addition to being familiar with the methods, potential users should also be able to conceptualise the problem.

The issue of spatio-temporal relationsips is demonstrated by means of the most frequent example – proximity. A square grid is often used as a simple but effective representation of proximity. Such an approach, however, suffers from rasterisation problems, when two objects in different neighbouring cells may in fact be closer than objects in the same cell. The rasterisation effects are partially mitigated in this study by using kernel densities to represent individual objects as a continuous surface. Still, the results depend on the raster resolution and the kernel bandwidth, while the approach leads to information loss. Other ways to represent the spatial neighbourhood exist, e.g. buffer zones around the features of interest [Karasová et al., 2005] or neigh-

bourhood graphs [Ester et al., 1997], which are more flexible, but computationally more demanding. Considering the suitability and benefits of other representations may be an interesting topic for further research.

As the methods presented here are data-driven, the reliability of the results depends on the quality of the input datasets. On the other hand, the nature of the phenomenon being studied allows some degree of generalisation to provide an overall view into the relationships behind it. The user should therefore consider the issue of data quality and be aware of the problems that may occur.

The insufficient positional accuracy of the incident dataset became clear during this research [Krisp et al., 2008]; this restricts the practical utilisation of the findings. The importance of the quality of incident records for further use is recognised by the fire & rescue services in Finland and special attention is being paid to data collection nowadays. The coordinates of the incident location are inserted into the database via an electronic report completed by the mission commander clicking a mouse on the corresponding place in the map. This process should ensure the highest possible accuracy. However, as the commander's main responsibility is extinguishing the fire, there may be some doubt as to the precision of the coordinates of incidents, which may not correspond to incident addresses. The responsible authorities should therefore consider using automatic methods for geospatial data collection.

The supplementary datasets used to provide background information must also be carefully selected. The resolution and spatio-temporal compatibility need to be considered with respect to the purpose of the analysis. This research illustrates the situation: the goal is to discover spatio-temporal relations existing in the data; however, the temporal variations of explanatory background variables are unknown, as, for example, population density data are based on the permanent addresses of the inhabitants. A detailed and temporally varying population model is necessary to observe unbiased patterns. Although also needful in other applications, such a detailed data source is usually not available directly, and its construction is not straightforward [Ahola et al., 2007].

Additional uncertainty emerges with data processing. The analysis involves splitting the datasets into various categories, which often have vague boundaries between them. The classification of the building types can serve as an example: where do hotels belong, in the housing, leisure, or work category? Although we do not expect a major influence of data classification on the results, this issue is postponed for further

analysis.

The reliability of the results is a complex problem the user has to cope with. The only solution to it is to consider the quality of the source data and understand the principles of the processing of the data. This study aims to help the user in assessing the reliability of the results by providing guidelines to the analysis process.

The estimate of a risk usually consists of the probability of its occurrence and the quantification of unwanted consequences based on particular scenarios. This study focuses on the former, analysing the spatio-temporal distribution of domestic fires. The severity of the impact of the fire, which may be independent of the probability of its occurrence, is also of interest to fire & rescue organisations for modelling the risk. The methods presented here could also be applied to identify factors influencing the consequences of fires, supposing that detailed fire statistics, including the costs of fires, are available as a source of information. For example, the weather conditions and construction materials used in a building, which may influence the speed at which the fire spreads, the ability of the inhabitants to react in emergency situations that may concern children, invalids, and elderly people, or the travel time needed for fire brigades to reach the incident location may be analysed. Risk modelling is, however, a complex problem that requires the consideration of more aspects than a probabilistic approach can provide [Sarewitz et al., 2003].

# Chapter 10

# Conclusions

The study concerns the process of discovering relationships from geospatial data with the aim of supporting the development of a fire risk model. It presents different approaches to the problem, with special attention being paid to an appropriate conceptualisation. As each of the methods applied originates in a different scientific field, it offers new viewpoints on the relationships being discovered, while posing different requirements for the user. The study documents all of these aspects to provide guidelines to the process for its future users. It also points out what can go wrong in the process, from conceptualisation and data processing to the interpretation of the results. In this way the study helps the users to find the most suitable approach, bearing in mind their requirements and skills.

## 10.1 Conceptualisation

One of the main challenges identified in this research is to find a suitable conceptualisation of the problem before the selected methods can be used. It refers to the way in which the geographical phenomenon being studied is presented in the form of appropriate data models to reflect the relationships of interest. The conceptualisation thus links the user's domain view of the problem to the actual analysis. A detailed analysis, which aims to support the understanding of the phenomenon being studied, requires a careful conceptualisation which integrates the expert knowledge into the model. Close cooperation between the domain and GI experts is necessary in order to perform the analysis in a way which satisfies the user's requirements. The

domain experts, with their knowledge and experience in their specific field, are the most important source of information. A GI scientist, on the other hand, provides his knowledge and tools to process geospatial data. He acts as a connection node between the domain of information science and the application field.

Conceptualisation is a complex process of interaction between the parties involved, so that the GI expert is aware of the user's needs and the user understands the potential and constraints of the analysis. At the beginning, it is useful to define the questions to be answered by the analysis. It may include a description of the behaviour of the phenomenon being studied over space and time. A preliminary analysis is performed in close cooperation with the domain experts using basic methods such as kernel densities and histograms. This step helps to define the spatial and temporal resolutions of the analysis with regard to the application in question.

The next step includes the determination of the physical extent of the analysis and the additional characteristics to be studied. The requirements for the additional datasets needed for the analysis can be defined and the datasets can be chosen accordingly.

Further, it is necessary to define the scope of the analysis with respect to the application and to decide about the type of patterns and relationships which are to be explored. These can be spatial or spatio-temporal, univariate, bivariate, or multivariate. Their characterisation can be qualitative or quantitative, local or global. If the user is interested in spatial patterns, he should be aware that these may change with time. The user might decide to analyse bivariate relationships, but he should know that the phenomenon being studied can be more complex and multivariate analysis may provide a more realistic description. In some cases a qualitative description of the phenomenon may be sufficient; other applications may require a more detailed quantitative approach or modelling. The user should know whether any local variations that may ensue are meaningful for him, or if he is interested in a global approach. The user should also be aware of the spatial and temporal scales he is interested in to study the global or local patterns. All these aspects must be considered when taking the purpose of the analysis into account.

The next step in the conceptualisation phase includes the required data preprocessing, conversions, and manipulation so that the method selected is technically capable of being applied. For a spatial analysis this step may include spatial transformations and the calculations of distances and densities or map layer overlays.

Chapter 8 demonstrates the influence of conceptualisation on the results. Two different approaches are used, each of which gives a different message to the user. Special attention must therefore be paid to this step, as the success of the analysis presumes an understanding of the user's requirements by the GI specialist as well as an insight into the analysis on the part of the user.

The study presents different approaches to conceptualisation, according to the method applied. Kernel density is used to acquire information about the spatial distribution of the phenomenon being studied. The advantage of the kernel surfaces, compared to point data, is in providing a more effective insight into the spatial distribution of the data, while allowing the level of detail preserved to be controlled by the user by means of the possibility of changing the kernel bandwidth. In this way, the kernel density also helps the user to define a suitable spatial scale for the analysis. With regard to the application and its purpose and scale, a kernel bandwidth of 200 m is found to be suitable in this study.

In order also to include the time dimension into the analysis, a preliminary analysis is performed in order to define a suitable temporal scale. As the most significant changes in the density of domestic fires occur between different perionds of the day, the three main temporal categories are selected to be analysed separately.

Additional datasets describing the factors influencing the distribution of domestic fires to be studied are selected with regard to the spatial resolution of the analysis. However, models that also reflect the temporal changes in these factors are not available. The datasets are associated according to the spatial location using a map overlay or distance calculations. The pre-processing may also require the re-classification or categorisation of the data. The user should be aware of the impact of such steps on the quality of the results.

## 10.2 Methods

The study brings an alternative viewpoint on the process of knowledge discovery. Instead of keeping the distinction between data mining and exploratory and statistical analysis, the study considers the concept of data mining and knowledge discovery as an umbrella concept, where the goal is of interest. The methods applied vary from visualisations, through statistical approaches, to computational methods originating in the field of artificial intelligence. The study proves their usefulness in discovering

spatio-temporal relationships, regardless of the framework they belong to.

Each of the methods yields new insights into the analysis, which contribute to the ensuing advancement of the formation of knowledge. Thus, rather than a single method being applied to analyse the data, the study demonstrates that the use of several different methods may enhance the knowledge that is discovered. The study therefore fulfils the aim of the research.

Some of the methods may overlap in their findings. However, each of the methods imposes different requirements on the user as regards his background knowledge and skills. All of these aspects need to be considered carefully in order for the approach to the analysis which best satisfies the user's needs to be chosen. Such an approach may consist of several steps to be used in a sequence or in parallel, e.g. applying visual methods to get an initial insight into the relationships before using some of the selected methods to examine the potentially interesting relationships more closely and, eventually, quantification and modelling. The results may be compared and refined. The study documents the findings and thus provides guidelines for the analysis process for potential future users.

Table 10.1 summarises the results that each of the methods brings according to several criteria. The framework refers to the dimensionality of spatio-temporal patterns, considering one, two, or more non-spatial attributes. Further, the approach to the analysis in terms of using qualitative or quantitative measures is considered. The methods are also examined for their ability to recognise global or also local patterns. Finally, the availability of a spatial reference, as a map, is considered.

Visual methods enable univariate, bivariate, and multivariate relationships to be explored. They are suitable for most users, as they can be used without prior theoretical skills. More advanced methods, such as a spaceFill or a PCP, however, can be more difficult to interpret correctly. In particular, a detailed exploration of the multivariate relationships, which is based on the interactive selection and display of the patterns in several visualisations, requires training. As the visualisations presented in this study are connected to a map, this approach allows the exploration of the spatial and also spatio-temporal distribution of the patterns and identification of local variations in the relationships. The results are described qualitatively; the strength of the relationships can only be estimated visually. In general, visual methods offer a fast and interactive insight into the data and can be recommended as a first step in the analysis.

| Method | Framework | | | Approach | | Extent | | | Description |
|---|---|---|---|---|---|---|---|---|---|
| | Uni- | Bi-variate | Multi- | Qualit. | Quantit. | Global | Local | Map | |
| Visual methods | | | | | | | | | |
| histogram | x | | | x | | x | x | x | frequency distribution |
| scatter plot | | x | | x | | x | x | x | correlation |
| spaceFill | | x | | x | | x | x | x | correlation |
| bivariate map | x | x | | x | | x | x | x | spatial distribution |
| PCP | x | x | x | x | | x | x | x | correlation, data structure |
| Contingency tables | | | | | | | | | |
| $\chi^2$ | | x | | | x | x | | | correlation |
| Cramér's $V$ | | x | | | x | x | | | strength of correlation |
| Point pattern analysis | | | | | | | | | |
| density | x | | | x | | x | x | x | spatial distribution |
| $\hat{G}-$function | x | x | | x | | x | x | | clustering, regularity, correlation |
| fitted model | | | x | | x | x | | x | modelling the relationships |
| GWR | | | | | | | | | |
| model, diagnostics | | x | x | | x | x | x | x | modelling the relationships |
| SOM | | | | | | | | | |
| U-matrix | | | x | x | | x | x | | clustering |
| component planes | x | x | x | x | | x | x | | correlation, data structure |

Table 10.1: Summary of the aspects revealed by the methods applied.

Contingency tables enable the correlations in the data to be quantified. They are primarily designed for bivariate analysis. As the phenomenon being studied can be more complex, the user should be aware of the possibly misleading results of such an analysis. Contingency tables can be applied to categorical variables. If the data being analysed are not categorical in nature, suitable conceptualisation and categorisation, which may result in a loss of information, is necessary. As a purely statistical approach, contingency tables offer limited possibilities for visualisation and mapping, which makes the interpretation and presentation of the results difficult. However, the method is useful for evaluating and comparing the relationships between two variables.

Point pattern analysis represents a statistical approach that has been specifically developed for dealing with spatial data. It can be applied to data in the form of point patterns, which enables the original level of detail to be preserved. First-order effects, expressed as a kernel density, are useful for describing the univariate distribution of the phenomenon being studied and to identify the hotspots. An adjustable kernel bandwidth controls the level of detail to be preserved. Measures based on the nearest neighbour functions as second-order effects enable the univariate and bivariate spatial dependence to be investigated. Modelling the spatial point processes considers multivariate relationships and allows them to be quantified globally. Visual representations of the results, such as density maps or distance function plots, facilitate the interpretation and communication of the results. Methods of point pattern analysis can be recommended to those users who are familiar with spatial statistics for a deep investigation and modelling of the processes being studied.

GWR is an advanced regression method that allows the local analysis of relationships in spatial datasets. The approach taken by GWR to the problem of local dependence and autocorrelation is natural. The use of GWR for the investigation of spatial processes, however, requires prior knowledge of statistics from the user in order for them to understand the processes being modelled. The possibility of mapping the GWR results facilitates their interpretation and the exploration of the data. This technique should be applied if the local variations in the relationships are of interest.

The SOM represents an alternative to statistical approaches. It is an artificial neural network, which maps the multidimensional space onto a two-dimensional lattice, while preserving similarity patterns existing in the input data. The SOM can be regarded as a computational clustering algorithm, which offers a variety of ways

to visualise the results. A distance matrix allows the clusters to be identified. Their characteristics can be defined from component planes, which are also useful for finding correlations. In this way the SOM provides an insight into multivariate relationships, while also allowing local relationships to be investigated. The use of the SOM, as an approximation of an input space, necessarily results in a loss of information. The method thus provides a qualitative description of the data. As an unconventional method, the SOM can easily be misused. An understanding of the SOM algorithm is necessary for the user in order to ensure that its application answers the proper questions and meets the expectations of the user. The SOM can be used for the advanced visual analysis of data based on a computational algorithm. The possibility of mapping the results to the original space, which is currently not supported by the software used in this study, would make the SOM useful for the exploration of geographical data. It is recommended as an alternative for those users who are not familiar with statistical approaches.

## 10.3 Implications for risk modelling

The study brings new knowledge to the issue of risk modelling. The following recommendations can be made on the basis of the results of this research.

- **Temporal variations during the different perionds of the day should be considered in risk modelling.**

  The current application is based on static models, which do not take temporal variations into account. The results of this research indicate the importance of including the time dimension in the analysis. For example, the occurrence of domestic fires is related to the population density at night, while other factors, such as the density of workplaces, seem to be more important during the daytime. The most significant changes in the distribution of fires occur during the different periods of the day, with daytime, evenings, and nights identified as the most important categories. These categories should be considered by the fire & rescue services for detailed temporal analysis in the future.

- **A thorough analysis is necessary to provide an insight into the background processes.**

Since the background processes behind the distribution of domestic fires are complex, a thorough analysis is necessary in order to provide an insight into the phenomenon. It should not be limited to simple methods and bivariate relationships; rather, a careful multivariate analysis should be performed. Using different methods which unveil different aspects of the relationships in the data enhances the understanding of the phenomenon being studied.

- **Special attention should be paid to the quality of the data acquired.**

  The methods presented in this study are data-driven. The reliability of the results depends on the quality of the input data. The importance of the data quality issue should be recognised by the fire & rescue organisations. Automatic data acquisition methods should be considered in order to ensure the highest quality of the data records.

- **Analysis of the background processes should be performed continuously.**

  Spatial phenomena are not constant but change in space and time. This also applies to the distribution of domestic fires. In order to capture the changing properties, analyses unveiling the relationships behind the spatial processes should be performed continuously whenever new data are available. The fire & rescue services should be able to perform such analyses in order to support the development of reliable risk models. This study documents the analysis process from conceptualisation and data pre-processing to the interpretation of the results, while describing the strengths and weaknesses of each technique. In this way it provides comprehensible guidelines for the users from the application domain, who may not be familiar with these issues.

- **An application integrating different suitable methods should be developed to facilitate the analysis.**

  A continuous analysis to be performed on the part of the user requires an application in which the methods selected are implemented via a user-friendly interface. The suitability of the methods should consider both the data model and the user's skills. Guidelines to the analysis in the form of detailed documentation of the methods should be provided in order to enable the methods to be used correctly.

## 10.4 Future research

As demonstrated by means of the case studies, the application of methods originating in different fields often faces problems of insufficient software support suitable for geospatial data. A connection to current GI software is also lacking. It would be desirable also to have powerful data mining capabilities within GI software packages specifically designed to handle spatio-temporal data. The findings documented within this study can support the specification of the requirements for a user-friendly geospatial data mining toolbox.

Naturally, the design of such a data mining toolbox is application-dependent. It entails a long process of interactive cooperation with the domain experts to specify their requirements and abilities. On the basis of the research presented here, however, general remarks can be made. A data mining toolbox should consider all the aforementioned aspects of mining geospatial data. In particular, it should offer a variety of distinct methods bringing new viewpoints to the analysis and thus realising the potential to reveal different aspects of relationships in the data. The collection should also consider the capabilities of the potential user to apply the method successfully and, besides sophisticated methods, also include powerful techniques requiring less background knowledge from the user.

This study demonstrates a need for an integrated toolbox for fire & rescue services. Future research should focus on the development of such a toolbox, offering several different methods via a single user interface. The methods should not be limited to those presented in this study, but the research could continue by analysing the suitability of other techniques to discover spatio-temporal relationships. Special attention should be paid to the development of the methods selected to provide a spatial and temporal reference. The toolbox should also allow for flexible data transformations and the comparable visualisation of the results. The possibility of implementing an 'analysis builder' with a user-friendly graphical interface to design the analysis process to be performed might be useful.

The results of this study motivate further research for fire & rescue organisations in Finland: RiskGIS is an ongoing project (2008–2010), in which Helsinki University of Technology and Inplace Solutions Ltd are developing a GIS application, which includes the methods studied in this research. The project is funded by the Finnish Funding Agency for Technology and Innovation (Tekes) in its Safety and Security

programme and the Internal Security ICT Agency. The National Emergency Supply Agency, the Emergency Services College and the Finnish Defence Forces are following and advising the work. The Fire and Rescue Foundation funded this research since 2005 until the project started. Within this project, a set of methods is identified as useful for the risk analysis process and a prototype application including data mining as a support for risk allocation is implemented on the basis of ArcGIS Objects and Java programming. In this way, the potential of the past incident records stored in the databases can be interactively exploited by end users to support their knowledge for the risk modelling. This research thus has a practical impact in the application domain, where it has lived up to its expectations.

# References

[Ackoff, 1989] Ackoff, R. L., 1989. From Data to Wisdom. Journal of Applied Systems Analysis 16 (1989): 3–9.

[Akaike, 1974] Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19 (6): 716–723.

[Agresti, 2002] Agresti, A., 2002. Categorical Data Analysis John Wiley & Sons, Inc, Hoboken, New Jersey.

[Ahola et al., 2007] Ahola, T., Virrantaus, K., Krisp, J. M. and Hunter, G. H., 2007. A spatio-temporal population model to support risk assessment and damage analysis for decision making. International Journal of Geographical Information Science, 21(8): 935–953.

[Andrews et al., 2003] Andrews, P. L., Loftsgaarden, D. O. and Bradshaw, L.S, 2003. Evaluation of fire danger rating indexes using logistic regression and percentile analysis. International Journal of Wildland Fire, 12 (2): 213–226.

[Ayyub, 2003] Ayyub, B. M., 2003. Risk Analysis in Engineering and Economics. Chapman&Hall/CRC, Boca Raton, Florida.

[Baddeley, 2008] Baddeley, A., 2008. Analysing spatial point patterns in R. CSIRO workshop notes [online]. Available from: http://www.csiro.au/files/files/pn0y.pdf [Accessed 1 April 2008].

[Baddeley and Turner, 2005] Baddeley, A. and Turner, R., 2005. Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software 12 (6): 1–42.

[Bailey and Gatrell, 1995] Bailey, T. C. and Gatrell, A. C., 1995. Interactive Spatial Data Analysis. Longman, Harlow.

[Bellinger et al., 2004] Bellinger, G., Castro, D. and Mills, A., 2004. Data, Information, Knowledge, and Wisdom. Available online: http://www.systems-thinking.org/dikw/dikw.htm (accessed 1 September 2009).

[Bishop et al., 1975] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge.

[Bowman and Azzalini,1997] Bowman, A. W. and Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. Oxford University Press, New York.

[Chawla et al., 2001] Chawla, S., Shekhar, S. and Ozesmi, U., 2001. Modelling spatial dependencies for mining geospatial data. In: Miller, H. J. and Han, J. (eds.) Geographic Data Mining and Knowledge Discovery. Taylor & Francis, London.

[Cochran, 1954] Cochran, W.G., 1954. Some methods for strengthening the common $\chi^2$ tests. Biometrics 10: 417–451.

[Cova, 1999] Cova, T. J., 1999. GIS in emergency management. In: Longley, P., Goodchild, M., Maguire, D., Rhind, D. (eds.) Geographic Information Systems, Volume 2. John Wiley & Sons, Inc., New York, pp. 845–858.

[Demšar, 2006] Demšar, U., 2006. Data mining of geospatial data: combining visual and automatic methods. Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden.

[Demšar, 2007] Demšar, U., 2007. Knowledge discovery in environmental sciences: visual and automatic data mining for radon problems in groundwater. Transactions in GIS, 11:255-281.

[Demšar et al., 2006] Demšar, U., Křemenová, O. and Krisp, J., 2006. Exploring geographical data with spatio-visual data mining. In: Kainz, W., Riedl, A. and Elmes, G. (eds.) Spatial Data Handling - Status Quo and Progress, Proceedings

of the 12th International Symposium on Spatial Data Handling, Vienna, Austria, pp. 149–166.

[Demšar et al., 2008] Demšar, U., Fotheringham A.S. and Charlton M., 2008. Exploring the spatio-temporal dynamics of geographical processes with Geographically Weighted Regression and Geovisual Analytics. Information Visualization, 7: 181–197.

[DiBiase, 1990] DiBiase, D., 1990. Visualization in the Earth Sciences. Earth and Mineral Sciences, 59(2): 13–18.

[Diggle, 2003] Diggle, P. J., 2003. Statistical Analysis of Spatial Point Patterns. 2nd ed. Arnold, London.

[Edsall, 2003] Edsall, R. M., 2003. The parallel coordinate plot in action: design and use for geographic visualization. Computational Statistics & Data Analysis 43: 605–619.

[Emergency Services College, 2009] Emergency Services College, 2009. Description available online: Statistics system of Finnish rescue services (PRONTO). http://prontonet.fi (accessed 6 November 2009).

[Ester et al., 1997] Ester, M., Kriegel, H.-P. and Sander, J., 1997. Spatial Data Mining: A Database Approach. In: Proceedings of the 5th International Symposium on Large Spatial Databases, SDD 1997, Berlin, Germany.

[Estivill-Castro and Lee, 2001] Estivill-Castro, V. and Lee, I., 2001. Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-refferenced Crime Data. In Proceedings of 6th International Conference on Geocomputation, Brisbane, Australia.

[Everitt, 1992] Everitt, B.S., 1992. The Analysis of Contingency Tables. Chapman & Hall/CRC, Boca Raton, Florida.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Sharpio, G. and Smith, P., 1996. From Data Mining to Knowledge Discovery: An Overview. In: Fayyad, U., Piatetsky-Sharpio, G., Smith, P. and Ulthurusamy, R. (eds.), Advances in Knowledge Discovery and Data Mining. Cambridge, MA: MIT Press: 1–34.

[Fayyad et al., 2002] Fayyad, U., and Grinstein, G. G., 2002. Introduction. In: Fayyad, U., Grinstein, G. G. and Wierse, A. (eds), Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, San Francisco.

[Fotheringham, 2009] Fotheringham, S. A., 2009. Geographically Weighted Regression. In: Fotheringham, S. A. and Rogerson, P. A. (eds.), The SAGE Handbook of Spatial Analysis. SAGE Publications Ltd, London: 243–254.

[Fotheringham et al., 2002] Fotheringham, S. A., Brunsdon, C., and Charlton, M. E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.

[Fotheringham et al., 2000] Fotheringham, S. A., Brunsdon, C., and Charlton, M. E., 2000. Quantitative geography: perspectives on spatial data analysis. SAGE Publications Ltd, London.

[Gahegan et al., 2002] Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F., 2002. Introducing GeoVISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. Computers, Environment and Urban Systems 26: 267–292.

[Godschalk, 1991] Godschalk, D., 1991. Disaster Mitigation and Hazard Management. In: Drabek, T. E. and Hoetmer G. J. (eds.) Emergency Management: Principles and Practice for Local Govermnent. International City Management Association, Washington, DC.

[Grinstein and Ward, 2002] Grinstein, G. G. and Ward, M. O., 2002. Introduction to Data Visualization. In: Fayyad, U., Grinstein, G. G. and Wierse, A. (eds), Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, San Francisco.

[Guo et al., 2005] Guo, D., Gahegan, M., MacEachren, A. M. and Zhou, B., 2005. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. Cartography and Geographic Information Science, 32(2): 113–132.

[Han et al., 2001] Han, J., Kamber, M. and Tung, A. K. H., 2001. Spatial clustering methods in data mining. In: Miller, H. J. and Han, J. (eds.) Geographic Data Mining and Knowledge Discovery. Taylor & Francis, London.

[Hand et al., 2001] Hand, D., Mannila, H. and Smyth, P., 2001. Principles of Data Mining. The MIT Press, Cambridge, Massachusetts.

[Helokunnas, 1995] Helokunnas T., 1995. Object-Oriented Approaches Applied to GIS Development. Acta Polytechnica Scandinavica, Mathematics and computing in engineering series No. 75, 1995.

[Hoetmer, 1991] Hoetmer, G. J, 1991. Introduction. In: Drabek, T. E. and Hoetmer G. J. (eds.) Emergency Management: Principles and Practice for Local Govermnent. International City Management Association, Washington, DC.

[Hope, 1968] Hope, A. C. A., 1968. A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society Series B, 30(3), 582–598.

[Ihamäki, 1997] Ihamäki, V-P., 1997. Paikkatietojärjestelmien (GIS) käyttö palo- ja pelastustoimen yhteistoiminnan suunnittelussa (Geographic information systems in planning cooperation between fire and rescue services). Pro Gradu Thesis, Helsinki University, Helsinki, Finland.

[Inselberg, 2002] Inselberg, A., 2002. Visualization and data mining of high-dimensional data. Chemometrics and Intelligent Laboratory Systems, 60: 147–159.

[Jashapara, 2004] Jashapara, A., 2004. Knowledge Management: An Integrated Approach. FT Prentice Hall, Pearson Education Limited, Harlow.

[Jenks, 1967] Jenks, G. F., 1967. The Data Model Concept in Statistical Mapping. International Yearbook of Cartography, 7: 186–190.

[Jiang and Harrie, 2004] Jiang, B. and Harrie, L., 2004. Selection of streets from a network using self-organizing maps. Transactions in GIS, 8: 335–350.

[Jolliffe, 2002] Jolliffe, I. T., 2002. Principal Component Analysis. 2nd edition, Springer, New York.

[Karasová et al., 2005] Karasová, V., Krisp, J.M. and Virrantaus K., 2005. Application of spatial association rules for development of a risk model for fire and rescue services. In: Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS), Stockholm, Sweden, pp. 183–194.

[Kelly and Ripley, 1976] Kelly, F. P. and Ripley, B. D., 1976. A note on Strauss's model for clustering. Biometrika, 63: 357–360.

[Keim, 2001] Keim D. A., 2001. Visual exploration of large data sets. Communications on ACM, 44(8): 38–44.

[Kiviluoto, 1996] Kiviluoto, K., 1996. Topology Preservation in Self-Organizing Maps. In: Proceedings of International Conference on Neural Networks (ICNN), Washington, DC, USA.

[Kohonen, 1997] Kohonen, T., 1997. Self-Organizing Maps. 2nd edition, Springer Verlag, Berlin-Heidelberg.

[Koperski and Han, 1995] Koperski, K. and Han, J., 1995. Discovery of Spatial Association Rules in Geographic Information Databases. In: Proceedings of 4th International Symposium on Large Spatial Databases, Portland, Maine, USA.

[Koua and Kraak, 2004] Koua, E. L. and Kraak, M-J., 2004. Alternative visualization of large geospatial datasets. The Cartographic Journal, 41: 217–228.

[Kraak and Ormeling, 2003] Kraak M.-J. and Ormeling F., 2003. Cartography, Visualization of Geospatial Data. 2nd edition, Prentice Hall, Harlow.

[Krisp, 2008] Krisp, J., 2008. Visualizing Population Density for Fire & Rescue services - An Application for Mobile Phone Location Data?, In: Mobile positioning data in geography and planning: data, analyses and applications, International workshop Social Positioning Method (SPM) 2008, Tartu, Estonia, pages pending.

[Krisp and Karasová, 2005] Krisp, J.M. and Karasová, V., 2005. The relation between population density and fire & rescue service incidents in urban areas.

In: Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS), Stockholm, Sweden, pp. 237–246.

[Krisp et al., 2005] Krisp, J., Virrantaus, K. and Jolma, A., 2005. Using explorative spatial analysis methods in a GIS to improve fire and rescue services, In: Oosterom, P., Zlatanova, S. and Fendel, E. M. (eds.) Geo-information for Disaster Management, Springer Berlin, pp. 1282–1296.

[Krisp et al., 2008] Krisp, J., Špatenková, O. and Ahola, T., 2008. Visual error detection in geocoded point data. Workshop of ICA Commission on Visualization: From Geovisualization toward geovisual analytics, Helsinki, Finland.

[Krisp and Špatenková, 2009] Krisp, J. and Špatenková, O., 2009. Kernel density estimations and their application in visualizing mission density for fire & rescue services. In: Proceedings of Joint Symposium of ICA Working Group on Cartography in Early Warning and Crisis Management and JBGIS Geo-information for Disaster Management - Cartography and Geoinformatics for Early Warning and Emergency Management: Towards Better Solutions, Prague, Czech Republic.

[Lonka, 1999] Lonka, H., 1999. Risk Assessment Procedures used in the field of civil protection and rescue services in different European Union countries and Norway. Helsinki: SYKE.

[MacEachren and Ganter, 1990] MacEachren, A. M. and Ganter, J. H., 1990. A pattern identification approach to cartographic visualization. Cartographica, 27(2): 64–81.

[MacEachren and Kraak, 2001] MacEachren, A. M. and Kraak, M-J., 2001. Research Challenges in Geovisualization. Cartography and Geographic Information Science, 28: 3–12.

[MacEachren et al., 2003] MacEachren, A. M., Dai, X., Hardisty, F., Guo, D. and Lengerich, G., 2003. Exploring High-D Spaces with Multiform Matrices and Small Multiples. In: Proceedings of the International Symposium on Information Visualization. Seattle, pp. 31–38.

[Maier, 2004] Maier, R., 2004. Knowledge Management Systems: Information and Communication Technologies for Knowledge Management, 2nd edition. Springer-Verlag Berlin Heidelberg.

[Malerba et al., 2001] Malerba, D., Esposito, F. and Lisi, F. A., 2001. Mining Spatial Association Rules in Census Data. In: Proceedings of the Joint Conferences on New Techniques and Technologies for Statistics and Exchange of Technology and Know-how, Crete, Greece.

[Mannila, 2002] Mannila, H., 2002. Local and Global Methods in Data Mining: Basic Techniques and Open Problems. In: Proceedings of 29th International Colloquium on Automata, Languages and Programming, Lecture Notes on Computer Science, 2380, Springer-Verlag pp. 57–68.

[Mattfeldt et al., 2007] Mattfeldt, T., Eckel, S., Fleischer, F. and Schmidt, V., 2007. Statistical modelling of the geometry of planar sections of prostatic capillaries on the basis of stationary Strauss hard-core processes. Journal of Microscopy 228 (3): 272–281.

[Miles and Huberman, 1994] Miles, M. B. and Huberman, A. M., 1994. Qualitative Data Analysis. SAGE Publications Inc., Thousand Oaks.

[Miller and Han, 2001] Miller, J. H. and Han, J., 2001. Geographic Data Mining and Knowledge Discovery: an overview. In: Miller, H. J. and Han, J. (eds.) Geographic Data Mining and Knowledge Discovery. Taylor & Francis, London.

[Modarres, 2006] Modarres, M., 2006. Risk Analysis in Engineering: Techniques, Tools, and Trends. CRC Press, Taylor & Francis Group, London.

[Molarius et al., 2009] Molarius, R., Rantanen, H., Huovila, H., Korpi, J., Yliaho, J., Wessberg, N., Virrantaus, K., and Rouhiainen, V., 2009. Assuring the information flow from accident sites to decision makers - a Finnish case study. In: Proceedings of the First International Conference on Disaster Management and Human Health Risk: Reducing Risk, Improving Outcomes, New Forest, UK.

[Mäkelä and Virrantaus, 2008] Mäkelä, J. and Virrantaus, K. 2008. Shared situational awareness in civilian crisis management. Extended abstract. Proceedings of GIScience 2008. Park City, Utah, USA.

[O'Sullivan and Unwin, 2003] O'Sullivan, D. and Unvin, D. J., 2003. Geographic Information Analysis. Wiley, Hoboken.

[Pearson, 1904] Pearson, K., 1904. On the Theory of Contingency and its Relation to Association and Normal Correlation. Drapers Company research memoirs: Biometric series I.

[Preisler et al., 2004] Preisler, H. K., Brillinger, D. R., Burgan, R. E. and Benoit, J. W., 2004. Probability based models for estimating wildfire risk. International Journal of Wildland Fire, 13 (2): 133–142.

[Rescue services in Finland, 2006] Rescue services in Finland, 2006. Available online: http://www.pelastustoimi.fi/en/in-brief/ (accessed 1 October 2007).

[Ripley, 1976] Ripley, B. D., 1976. The second-order analysis of stationary point processes. Journal of Applied Probability, 13, 255–266.

[Saarinen and Hämäläinen, 2004] Saarinen, E. and Hämäläinen, R. P., 2004. Systems intelligence: Connecting engineering thinking with human sensitivity. In: Saarinen, E. and Hämäläinen, R. P. (eds) Systems intelligence: Discovering a hidden competence in human action and organisational life. Systems Analysis Laboratory Rearch Reports A88, Helsinki University of Technology, Espoo, Finland, pp. 9–38. Available online: www.systemsintelligence.hut.fi.

[Santos and Amaral, 2004] Santos, M. Y. and Amaral, L. A., 2004. Mining geo-referenced data with qualitative spatial reasoning strategies. Computers & Graphics, 28: 371–379.

[Sarewitz et al., 2003] Sarewitz, D., Pielke, R., Keykhah, M., 2003. Vulnerability and Risk: Some Thoughts from a Political and Policy Perspective. Risk Analysis, 23/4, 805–810.

[Schneiderman and Plaisant, 2005] Schneiderman, B. and Plaisant, C., 2005. Designing the User Interface, Strategies for Effective Human-Computer Interaction, 4th edition. Pearson Addison Wesley, Boston, USA.

[Seppänen and Virrantaus, 2009] Seppänen, H. and Virrantaus, K., 2009. The Role of GIS Methods in Crisis and Disaster Management. To appear in International Journal of Digital Earth in 2009.

[Shekhar and Chawla, 2003] Shekhar, S. and Chawla, S., 2003. Spatial Databases - A Tour, 2nd edition, ch. 7: Introduction to Spatial Data Mining, 182–226. Prentice Hall, Pearson Education Inc., Upper Saddle River, New Jersey, USA.

[Silipo, 2003] Silipo, R., 2003. Neural Networks. In: Berthold M and Hand DJ (eds), Intelligent Data Analysis, 2nd edition. Springer Verlag, Berlin-Heidelberg, 269–320.

[Silverman, 1986] Silverman B. W., 1986. Density estimations for statistics and data analysis. Chapman and Hall, London.

[StatSoft, 2008] StatSoft, 2008. Electronic Statistics Textbook. StatSoft, Inc., Tulsa, USA. Available online: http://www.statsoft.com/textbook/Stathome.html (accessed 1 September 2008).

[Strauss, 1975] Strauss, D. J., 1975. A model for clustering. Biometrika, 62: 467–475.

[Špatenková and Krisp, 2007] Špatenková, O. and Krisp, J., 2007. The use of contingency tables to value variables for spatial models. In: Proceedings of the 5th International Symposium on Spatial Data Quality, Enschede, the Netherlands.

[Špatenková et al., 2007] Špatenková, O., Demšar, U. and Krisp, J., 2007. Self-Organising Maps for exploration of spatio-temporal emergency response data. In: Proceedings of Geocomputation 2007, Maynooth, Ireland.

[Špatenková and Stein, 2009] Špatenková, O. and Stein, A., 2009. Identifying factors of influence in the spatial distribution of domestic fires. To appear in International Journal of Geographical Information Science.

[Statistics Finland, 2006] Statistics Finland, 2006. Grid database. Description available online: http://tilastokeskus.fi/tup/ruututietokanta (accessed 6 November 2009).

[Takatsuka and Gahegan, 2002] Takatsuka, M. and Gahegan, M., 2002. GeoVISTA Studio: a codeless visual programming environment for geoscientific data analysis and visualization. Computers & Geosciences 28: 1131–1144.

[Tillander, 2004] Tillander, K., 2004. Utilisation of statistics to assess fire risks in buildings. VTT Technical Research Centre of Finland.

[Unwin and Unwin, 1998] Unwin, A. R. and Unwin, D., 1998. Exploratory spatial data analysis with local statistics. The statistician 47: 415–423.

[Vesanto, 1999] Vesanto, J., 1999. SOM-based data visualization methods. Intelligent Data Analysis, 3: 111–126.

[Vesanto et al., 2000] Vesanto J., Himberg J., Alhoniemi, E., Parhankangas, J., 2000. SOM Toolbox for Matlab 5. Report A57. Libella Oy, Espoo, Finland.

[Virrantaus, 2009] Virrantaus, K., 2009. Use of Spatial Data Mining in Risk and Vulnerability Assessment for Crisis and Disaster Management – A System Intelligent Approach. Keynote speech. 17th annual GIS Research UK (GISRUK) conference, Durham, UK.

[Worboys and Duckham, 2004] Worboys, M. and Duckham, M., 2004. GIS: a computing perspective, 2nd edition. CRC Press LLC, Boca Raton, Florida.

[YTV, 2003] YTV - Helsinki Metropolitan Area Council, 2003. SeutuCD03. Description available online (in Finnish): http://www.ytv.fi/FIN/seutu_ymparistotietoja/tietoaineistot/seutucd (accessed 6 November 2009).

[Zhang and Goodchild, 2002] Zhang, J. and Goodchild, M. F., 2002. Uncertainty in Geographical Information. Taylor & Francis, London.